

Copyright
by
David Robert Beachum
2019

The Thesis Committee for David Robert Beachum
certifies that this is the approved version of the following thesis:

**Methods for Assessing the Safety of Autonomous
Vehicles**

APPROVED BY

SUPERVISING COMMITTEE:

Raul G. Longoria, Supervisor

Junmin Wang

**Methods for Assessing the Safety of Autonomous
Vehicles**

by

David Robert Beachum

Thesis

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2019

Acknowledgments

My sincerest gratitude goes to my thesis supervisor, Dr. Raul Longoria, for his guidance, throughout both this project and my time at the University of Texas.

I would also like to thank Dr. Junmin Wang for meeting to discuss the safety of autonomous vehicles, and for his interest in this project.

Lastly, thank you to my parents, Matt and Lalla Beachum, for their support and inspiration, and to Lexie Hassien for her patience and encouragement.

Methods for Assessing the Safety of Autonomous Vehicles

David Robert Beachum, MSE
The University of Texas at Austin, 2019

Supervisor: Raul G. Longoria

While the widespread adoption of autonomous vehicles (AVs) has the potential to drastically reduce the rate of traffic collisions, failure to verify their safe operation may expose the public to unacceptable risks. Due to the low frequency of traffic fatalities, verifying AV safety statistically via on-road testing is likely to be cost- and time-prohibitive, driving the need for alternate methods. This thesis examines four potential methods to assess AV safety: simulation, Failure Modes and Effects Analysis (FMEA), Fault Tree Analysis (FTA), and Systems Theoretic Process Analysis (STPA). The findings show two methods with potential: simulation based on data recorded from real environments, and quantitative FTA combined with a secondary analysis of the vehicle's machine learning algorithms. However, both approaches require significant amounts of data which may be expensive to gather. Further research into the safety of machine learning algorithms and further developments in AV simulation technology are required in order to develop more cost-effective methods for assessing AV safety.

Table of Contents

Acknowledgments	iv
Abstract	v
List of Tables	viii
List of Figures	ix
Chapter 1. Introduction	1
1.1 The State of Autonomous Vehicles	2
1.2 The Importance of Assessing Safety	5
1.3 Summary	9
Chapter 2. Existing Methods for Assessing Safety	10
2.1 Testing	10
2.2 Simulation	13
2.3 Analytical Methods	14
2.3.1 Failure Modes and Effects Analysis	14
2.3.2 Fault Tree Analysis	17
2.3.3 Systems Theoretic Process Analysis	21
2.4 Summary	28
Chapter 3. Methodology	29
3.1 Simulation	29
3.1.1 Vehicle Model	30
3.1.2 Test Conditions	32
3.1.3 AV Controller	33
3.1.3.1 Failure Modes	34
3.1.3.2 Statistics	36

3.2 Analytical Methods	36
3.3 Summary	37
Chapter 4. Results	38
4.1 Simulation	38
4.2 Analytical Methods	43
4.2.1 Failure Modes and Effects Analysis	43
4.2.2 Fault Tree Analysis	47
4.2.3 Systems Theoretic Process Analysis	49
Chapter 5. Discussion	56
5.1 Simulation	56
5.2 Analytical Methods	60
5.2.1 Failure Modes and Effects Analysis	60
5.2.2 Fault Tree Analysis	63
5.2.3 Systems Theoretic Process Analysis	67
5.3 Assessing Safety of Machine Learning Systems	69
5.4 The Future of Regulation	71
Chapter 6. Conclusions and Future Work	76
6.1 Conclusions	76
6.2 Future Work	78
6.3 Summary	79
Bibliography	81

List of Tables

2.1	Disengagement report figures for top 10 developers by miles per disengagement.	11
2.2	Some common symbols used in FTA diagrams.	18
2.3	Accidents for the example APA system.	24
2.4	Hazards and safety constraints for the example APA system. .	24
2.5	Excerpt of Unsafe Control Actions (UCAs) for the example APA system.	27
3.1	Specifications for the simulated AV.	31
3.2	Open-loop turn sequence executed by the AV.	34
3.3	Failure modes applied to the simulated AV controller.	35
4.1	Collision rate and time taken to turn for different combinations of failure mode parameters.	41
4.2	Ratings used for the FMEA analysis.	44
4.3	Probability of basic, intermediate, and top-level failure events during the left turn maneuver.	47
4.4	Accidents for the left turn maneuver.	49
4.5	Hazards and safety constraints.	50
4.6	Unsafe Control Actions (UCAs) for the left turn maneuver. . .	52
4.7	Possible causes and recommended actions for UCA-2.	55

List of Figures

1.1	SAE classifications of autonomous driving systems.	3
2.1	Excerpt of an FMEA performed for an AV system by Tokody et al.	16
2.2	FTA performed by UTRC considering failure due to vehicular components.	19
2.3	FTA performed by UTRC considering failure due to infrastructure.	20
2.4	General model of a STAMP system.	22
2.5	Control structure for the example APA system.	25
3.1	Variables of interest in the vehicle model used in this simulation.	31
3.2	Layout of the simulated intersection.	32
4.1	Sample of a successful trial of the simulation.	39
4.2	Sample of trial of the simulation ending in collision caused by the sensor failing to detect an oncoming vehicle.	40
4.3	Interacting effects of sensor error in position and velocity measurements.	42
4.4	FMEA analysis for the vehicle completing a left turn.	45
4.5	FMEA analysis for the vehicle completing a left turn, continued.	46
4.6	Fault tree analysis performed for the left turn maneuver. . . .	48
4.7	Control structure for the AV system.	51

Chapter 1

Introduction

The prevalence and scale of automotive travel in modern society gives autonomous vehicles (AVs) tremendous capacity to prevent, or cause, fatalities. It is critical to verify the safety of these systems before they are put into public use, but many current methods for assessing AV safety are insufficient. Furthermore, the pressures felt by automotive manufacturers to be among the first to deliver self-driving cars to market, combined with federal regulation lagging behind innovation, may lead to AVs entering the market before their safety is fully verified. This thesis will examine methods that could be used to assess the safe operation of AVs.

The phrase “autonomous vehicle” can describe a wide range of vehicles, including aircraft, watercraft, and robots used in warehouses and other applications. However, in this work, “autonomous vehicles” will be used to refer specifically to passenger- or cargo-carrying ground vehicles used on public roads, commonly called self-driving or driverless cars.

In this thesis, Chapter 1 briefly introduces the history and current state of self-driving cars, and establishes the importance and difficulty of assessing their safety. Chapter 2 discusses five empirical and analytical methods that have been used to assess the safety of autonomous vehicles. The remainder of

the work describes how four of these methods can be implemented on a specific maneuver to evaluate each method’s ability to assess the safety of autonomous vehicles. Chapters 3, 4, and 5 present the methodology, results, and discussion of those analyses, respectively. Chapter 6 summarizes the findings of this thesis, and discusses potential future work in this field.

1.1 The State of Autonomous Vehicles

Although the idea of self-driving cars had been explored earlier in the form of radio-controlled [7] and guided path-tracking vehicles [8], the development of AVs as they are known today began in the 1980s, with a vision-guided vehicle developed by researchers at Bundeswehr University Munich [15] and the LIDAR-equipped Autonomous Land Vehicle (ALV) project developed at several American universities [29]. Research continued through the 1990s and 2000s, with notable works including Dean Pomerleau’s 1992 PhD thesis on the application of neural networks in autonomous vehicle navigation [52]. Efforts to make self-driving cars commercially available began without public announcement by Google in 2009, with the formation of the company that become Waymo in 2016 [53], with a number of other companies and automobile manufacturers beginning development on their own AV systems in the mid 2010s.

SAE International establishes a scale from 0 to 5 for categorizing the autonomy of an AV system [3], detailed in Figure 1.1. Level 0 systems may include warnings or momentary interventions, but no sustained control. Levels

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
Human driver monitors the driving environment						
0	No Automation	the full-time performance by the <i>human driver</i> of all aspects of the <i>dynamic driving task</i> , even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	Driver Assistance	the <i>driving mode</i> -specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	Human driver and system	Human driver	Human driver	Some driving modes
2	Partial Automation	the <i>driving mode</i> -specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	System	Human driver	Human driver	Some driving modes
Automated driving system ("system") monitors the driving environment						
3	Conditional Automation	the <i>driving mode</i> -specific performance by an <i>automated driving system</i> of all aspects of the dynamic driving task with the expectation that the <i>human driver</i> will respond appropriately to a <i>request to intervene</i>	System	System	Human driver	Some driving modes
4	High Automation	the <i>driving mode</i> -specific performance by an automated driving system of all aspects of the <i>dynamic driving task</i> , even if a <i>human driver</i> does not respond appropriately to a <i>request to intervene</i>	System	System	System	Some driving modes
5	Full Automation	the full-time performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> under all roadway and environmental conditions that can be managed by a <i>human driver</i>	System	System	System	All driving modes

Figure 1.1: SAE classifications of autonomous driving systems. Reproduced from SAE [3].

1 and 2 involve varying degrees of automatic assistance, but the driver must still constantly monitor the driving environment. At levels 3 and 4, the system monitors the driving environment under certain circumstances, although the driver must be ready to resume control. At level 5, or full autonomy, the system does not require human intervention under any driving modes.

In April 2017, Waymo began offering an autonomous taxi service to a small group of a few hundred early riders in several suburbs of Phoenix, AZ. In December 2018, this service launched commercially as Waymo One to the same limited group [35]. The AVs used in this program are among the most advanced in the world, arguably with Level 4 autonomy. However, they continue to struggle with certain tasks, such as merging onto highways and making unprotected left turns, and they are often criticized as driving too conservatively for their customers' satisfaction [18]. Among vehicles commercially available to the public, the Traffic Jam Pilot feature of the Audi A8 is currently the only AV system with Level 3 automation [14] (outside of the United States only). Level 2 systems such as Tesla's Autopilot, Volvo's Pilot Assist, Mercedes-Benz's Drive Pilot, and others are widely available to the public [27].

Forecasts for when fully autonomous (Level 5) vehicles will become widely available vary. One 2014 study aggressively predicted most automobile manufacturers will have fully autonomous vehicles for sale by 2020 [10]; timelines from eleven automobile manufacturers fall a few years later, with most planning on selling Level 3 systems in the very early 2020s, and Level 4 and

5 systems around 2025-2030 [69]. However, some claim many of these manufacturers are behind schedule, and point to issues encountered during testing - such as difficulties with unprotected left turns, or driving in inclement weather - to indicate AVs will arrive later than many think [60]. A 2018 study more conservatively predicts limited access to the affluent in the 2020s and 2030s, with Level 5 AV technology becoming widespread in the 2040s and 2050s [40].

1.2 The Importance of Assessing Safety

The prospect of AVs partially or wholly replacing conventional vehicles is attractive from the standpoint of safety. With approximately 1.25 million traffic fatalities recorded worldwide each year [5], and approximately 90%-93% of all accidents caused by human error [63], full automation could have the potential to save upwards of a million human lives annually. Additionally, widespread adoption of AVs could lead to fewer vehicles per capita, resulting in reduced traffic and parking congestion and reduced environmental impact, with each AV netting \$2,000-\$4,750 per year in societal benefit [19].

However, the potential of AVs to pervade and reshape our society also raises concerns of safety. Even in these early stages, the race to autonomy has proven fatal: on January 20, 2016, a Tesla Model S, operating under the Level 2 Autopilot feature, became the first AV to be involved in a fatal collision when it failed to brake on a collision course with a road sweeper vehicle, killing the driver of the Tesla [12]. The first fatal collision caused by a Level 3 system, and the first fatality to a person other than the operator of the AV, occurred on

March 18, 2018, when an AV operated by Uber failed to react to a pedestrian crossing a street at night in Tempe, Arizona [24]. In total, four fatalities have been attributed to AV technology: three by Tesla’s Autopilot, and one by an Uber vehicle. As AV technology becomes more widespread, one can expect the death toll to rise, and the safety of these technologies to enter the forefront of public discussion.

As of April 2019, AVs have been involved in 139 collisions in California alone [49], and they appear to be involved in accidents at a rate five times higher than human drivers [55]. Paradoxically, the vast majority of these accidents are caused by humans, not self-driving technology [32]. This does not absolve AVs, instead suggesting that AVs drive in a way that human drivers do not expect; that is, even when the AV as an individual component behaves safely, it may still introduce danger to the wider system. For example, AVs do not recognize many forms of communication used by human drivers, such as hand signals and eye contact.

Beyond the danger of traffic collisions, there are other societal risks associated with the widespread adoption of self-driving technology. Some researchers have raised concerns regarding the susceptibility of AV systems to cyberattacks [51]. Others have questioned how AVs and other artificial intelligence systems will fit into existing frameworks of legal liability, suggesting that if an AV is found to be even 1% responsible for an accident, the manufacturer may be required to pay 100% of the damages [33]. AV technology may present logistical challenges, with one researcher suggesting that in cer-

tain areas, it may be less expensive for AVs to cruise rather than park when not in use, potentially bringing city centers to gridlock [46]. And perhaps the most often discussed societal hazard of self-driving technology is its effects on unemployment, considering that, as of 2018, 4.8 million Americans (3.2% of the workforce) drive for a living [4].

Although the potential benefits of widespread use of AVs are clear, it is critical that the safety of these technologies be rigorously tested and systematically assessed. In conventional vehicles, a major mechanism for ensuring safety is government regulation. In the United States, the National Highway Traffic Safety Administration (NHTSA) issues Federal Motor Vehicle Safety Standards (FMVSS) which regulate nearly every aspect of a vehicle's form and function with the goal of ensuring safety [2]. FMVSS are particularly concerned with human interfaces, and tightly regulate features such as rear view mirrors and steering wheels - many of which would have no role in a fully-autonomous vehicle. The very idea of autonomous vehicles violates a range of FMVSS specifications, leading AV manufacturers to petition the NHTSA for interpretations of and exemptions from FMVSS, which have, by-and-large, been granted [54].

While the federal government has taken a very limited role in regulating AV technology, state governments have been somewhat more active, with 36 states passing laws or issuing executive orders on the matter [6]. However, some note that states seem to compete to have the most lenient regulations to attract the business of AV developers [25].

The result is that currently almost the entire burden of ensuring safety in AVs falls to the companies designing them, creating potentially dangerous conflicts of interest. In December of 2018, Waymo launched a limited-access self-driving taxi service with virtually no federal oversight [37], with other companies planning to launch their own services in the coming years [67]. As companies race to be among the first to put AVs to market, some companies may relax their own safety standards for an advantage in that race.

It is clear that if AVs are to reach their potential on a global scale, it is paramount that their safety be assessed and verified, but determining a threshold that is “safe enough” is not straightforward. One intuitive standard is that the safety of AVs is acceptable when the rate of collisions for AVs is lower than that of a human-driven vehicle. However, it is not clear whether all collisions should be included, or only those resulting in fatalities, personal injury, or property damage. Human bias further complicates the issue, with some evidence suggesting humans judge algorithms more harshly than they judge human operators, even when the algorithms perform better [16]. Other studies suggest human perception of technological risk is heavily biased by emotion [50]. Furthermore, highly reliable AV systems with less than full automation may cause their drivers to pay less attention by gaining their confidence, counter-intuitively increasing the rate of accidents [44] [58]. Human biases against autonomous technology, combined with dangers inherent to AVs with less than fully autonomy, may require that AV technology actually be far safer than conventional vehicles before they are determined

to be “safe enough” for the public. Although there is no clear and definitive answer to the question of defining a safety standard, the remainder of this work will discuss safety in terms of rate of collision of any type, compared to collision rate for human-driven vehicles.

1.3 Summary

The rapid development of AV technology in recent years is expected to continue, with fully autonomous vehicles available to the public within the next decade, and widespread adoption occurring in the decades after. The pressure on manufacturers to deliver AVs to market, combined with the limited approach taken by regulators, may result in unsafe vehicles reaching the public. Due to the prevalence of automotive travel in modern society, it is essential that the safety of AVs be verified rigorously.

The following chapter discusses existing methods that could be used to assess safety in AVs, and the remaining chapters of this thesis implement these methods and compares the results.

Chapter 2

Existing Methods for Assessing Safety

2.1 Testing

By far the most publicized method manufacturers have used to verify the safety of their AVs is by testing them on roads. Companies commonly report two metrics: total miles traveled, and miles traveled per disengagement. These metrics are popular largely because the state of California, a hotbed for AV testing and development, requires annual reports of both figures from all AV developers testing on public California roads [48].

Critics of these metrics argue that reducing the results of testing to a single number of miles removes all information regarding testing conditions. Although the number of miles driven is commonly reported, it is much less commonly reported what percent of these miles were driven on public roads versus private tracks, city roads versus highways, at what speeds, at what times of day, and during what weather conditions. Furthermore, the objective of tests is not clear, with some critics arguing it is impossible to know how many of these tests were done to gather data regarding real-world scenarios, and how many were done simply to accumulate miles for the sake of public perception [43].

Table 2.1: Disengagement report figures for top 10 developers by miles per disengagement, California only, Dec 2017 - Nov 2018 [26].

Developer	Miles	Miles/disengagement
Waymo	1,271,587	11,154.3
GM Cruise	447,621	5,204.9
Zoox	30,764	1,922.8
Nuro	24,680	1,028.3
Pony.AI	16,356	1,022.3
Nissan	5,473	210.5
Baidu	18,093	205.6
AIMotive	3,428	201.6
AutoX	22,710	190.8
Roadster.AI	7,539	175.3

A similar metric, miles per disengagement, measures the average distance traveled between incidents which cause the human driver to disengage the autonomous driving system and assume manual control. Again, very little information is reported regarding the human driver’s decision or the driving conditions at the time of these disengagements. Another shortcoming is that disengagements are a critical part of the learning process during AV development, and a lack of disengagements might indicate a lack of progress [43]. A company’s desire to report low disengagement rates to the public might incentivize overly-conservative, unproductive testing conditions. The fact that test results are reported using these overly-simplified metrics may be misleading and detrimental to public understanding of AV safety.

Beyond issues of public misperception, testing alone may not be viable for establishing safety. In 2013, American drivers caused just 77 reported in-

juries and 1.09 fatalities per 100 million miles driven [28]. Events this rare require extraordinarily large sample sizes to calculate probability with reasonable precision and accuracy. Researchers at RAND Corporation found that “autonomous vehicles would have to be driven hundreds of millions of miles and sometimes hundreds of billions of miles”, taking “tens and sometimes hundreds of years,” in order to statistically verify the safety of AVs [28] - too slow for even conservative estimates of when AV technology will become widespread. As of July 2018, Waymo, by far the frontrunner among AV developers in terms of miles travelled, claimed to have driven a total of 8 million miles, adding about 750,000 more every month [34]. The rate is likely to increase as fleet sizes increase, meaning Waymo and a select few others may very well accumulate hundreds of million miles over the next few decades. However, the prospect of repeating this process for every other manufacturer and for every new model will likely prove cost- and time-prohibitive.

Additionally, AVs present a particular challenge to road testing because the machine learning techniques used in decision making are much more difficult to verify correct operation in than conventional algorithms [42]. According to researchers at Technische Universität Darmstadt “current test concepts are not suitable for economically assessing the safety of a new system such as autonomous driving. Adhering to current test concepts would involve an economically unjustifiable overhead, and would result in an ‘approval-trap’ for autonomous driving.” [68]. Ultimately, alternative methods to road testing are needed to verify the safety of AVs before widespread use.

2.2 Simulation

Simulation provides another option for testing the reliability and efficacy of AV software. Simulation is most often performed using sensor data gathered from real environments [57], so that an AV can repeatedly attempt a scenario it encountered on the road, but a select few companies have developed software, such as Waymo’s Carcraft, that allows virtual vehicles to navigate fully simulated environments [41]. Simulation platforms capable of testing AVs in fully simulated environments are also offered by third-party companies such as Metamoto [45] and NVIDIA [47], which may significantly increase access to simulation for smaller AV manufacturers.

Simulation can eliminate many major drawbacks of road testing, such as issues of safety, liability, and regulation. It allows developers to isolate particularly dangerous scenarios for repeated testing with slightly differing parameters and conditions. Lastly, it can be carried out far more quickly than road testing. Waymo, for example, has logged 8 million miles on public roads, but 5 billion in simulation [34].

However, testing in simulation retains many of the other limitations of road testing. The same metric reported for road testing, miles driven, is reported for simulation, with the same underlying problems, although these figures can be even more misleading to the public, who cannot be expected to know the specifics of the simulation or gauge the value of a virtual mile against a mile on the road. Ultimately, the data from these tests is only as valuable as the simulation is accurate; elements that could result in a dangerous

environment might not be properly accounted for, and it is difficult to verify that the simulated environments are representative of actual environments.

2.3 Analytical Methods

A number of analytical methods have been developed and utilized to assess risk in any system, constituting the broad discipline of reliability engineering. Many of these methods have been applied to AV systems. This thesis will briefly consider two classical methods of risk assessment, Failure Modes and Effects Analysis (FMEA) and Fault Tree Analysis (FTA), before discussing a more recent method proposed for more complex dynamic systems, Systems Theoretic Process Analysis (STPA).

2.3.1 Failure Modes and Effects Analysis

Failure Modes and Effects Analysis (FMEA) is an early form of reliability engineering analysis initially documented by the United States military in the 1940s [64], and further developed and expanded over the following decades. Although the exact methodology varies, the core idea is to assess and minimize risks by identifying failure modes of each component in a system, and assigning each failure mode a Risk Priority Number (RPN), calculated as the product of that failure mode’s relative severity, probability, and detectability [61]. Each failure mode is then addressed with additional controls or redundancies to reduce or eliminate risk.

FMEA is one method that has been applied to AV systems in order to

assess their risk. Figure 2.1 shows an excerpt of an FMEA worksheet performed by Tokody et al. [62] as part of an article discussing the future impacts of AV systems. Note that the columns “Sever[e]”, “Occur”, and “Detect” provide ratings from 1-10 on the relative damage caused by each failure mode, the likelihood of that failure mode occurring, and the likelihood that the fault could be identified and detected before failure occurs, respectively.

Although this particular analysis is applied to an AV system at a high level, FMEA can be applied in other ways as well. For example, FMEA could be performed not for the entire system, but for a particular maneuver or function, such as automated parking, lane keeping, or turning at an intersection. FMEA has also been proposed to facilitate in an AV system’s decision making by updating the worksheet continuously onboard the vehicle using data from the environment, such that the AV’s software makes decisions that minimize risk [30].

Despite its widespread application across many fields, FMEA is subject to a number of limitations and criticisms. First, it is only able to identify known faults - that is, failure modes that the analysts performing the FMEA are able to foresee or have some amount of data for. Therefore, it may not be useful in identifying the potentially hazardous unforeseen interactions of a complex system, such as an AV in city traffic. It may also be less useful in emerging technologies, in which all failure modes have not yet occurred and there is limited data to draw from. As a bottom-up technique, FMEA has been described as having very little utility in safety analysis [39]. Additionally, in a

Number	Failure mode	Failure effect	Sever	Cause	Occur	Detect	RPN	Provisions
1	Turning off human intervention	Speeding irregular direction	10	Free provision over the system	10	10	1000	Restricted overwrite options
2	Hacker attack	Taking over the vehicle's control/incorrect data communication	10	System is not sufficiently encrypted	6	10	600	Testing
3	Faulty code	Speeding collision hazard	10	Human mistake	5	10	500	Multiple quality controls and tests
4	Weather anomalies	Change in braking distance data loss	10	–	5	10	500	Preparing the system for extreme weather
5	Wheeled vehicle signal loss	Positioning loss	10	Technical failure, shielding	4	9	360	In case of signal loss, change to manual control
6	Iron wheeled vehicle signal loss	Positioning loss	10	Technical failure, shielding	4	9	360	In case of signal loss, change to manual control
7	Distinguished vehicle handling	Distinguished vehicles are not integrated to the system, therefore increased accident risk	10	Incorrect handling of distinguished vehicle	7	6	420	Indicating distinguished vehicles to the system
8	Oversized vehicle handling	Oversized vehicle badly positioned delimitation thus creating a dangerous situation	10	Incorrect handling of oversized vehicle's specialties	4	8	320	Data handling of oversized vehicles

Figure 2.1: Excerpt of an FMEA performed for an AV system by Tokody et al. Reproduced from Tokody et al. [62]







1993 article, Warren Gilchrist argued that “though the model itself is of great use, the calculation of the RPN lacks a proper model as a base and thus is internally inconsistent and potentially misleading” [23]. In other words, the unweighted multiplication of subjective, ordinal scores for severity, occurrence, and detectability is arbitrary and may misguide efforts at reducing risks.

2.3.2 Fault Tree Analysis

The ratings given for severity, occurrence, and detectability during FMEA are commonly subjective estimates rated on a scale from one to ten, based on the experiences of the analysts, but using more meaningful numbers here can lead to more meaningful RPNs. Probabilistic Risk Assessment (PRA) is a field of reliability engineering that methodically quantifies the severity and probability of a failure mode and expresses their product as an expected loss. There are a variety of methods used within PRA to calculate failure severity and probability. This thesis will discuss Fault Tree Analysis (FTA), one such method for deriving the probability of system failure.

FTA was first developed in 1961 by Bell Telephone Laboratories during work on a contract studying a US Air Force missile launch control system, and subsequently gained popularity through the 1960s [38]. FTA provides an approach to determining the probability of system failure by constructing a logic tree of all possible basic failure events. If data is available for the probability of each basic failure event over a given length of time, the probability of system failure (or any intermediate category of failure) can be calculated

Table 2.2: Some common symbols used in FTA diagrams.

	AND gate: All input events are required to result in output event.
	OR gate: At least one input event is required to result in output event.
	Basic event.
	External event.
	Undeveloped event.
	Intermediate.

using Boolean logic. Events are connected by logic gates, which can be used to represent redundancy in a system. Some commonly used symbols are shown in Table 2.2.

FTA has been applied to AV systems, as shown in Figures 2.2 and 2.3 reproduced from a report prepared by Rowan University [65]. FTA has been criticized for its high cost of development when compared to simpler methods [37]. However, its ability to depict the relationships between failure events and synthesize probabilities for the entire system make it useful in complex systems with many possible failure events. As with all forms of quantitative risk analysis, FTA is limited by the quality of data available to produce accurate probabilities.

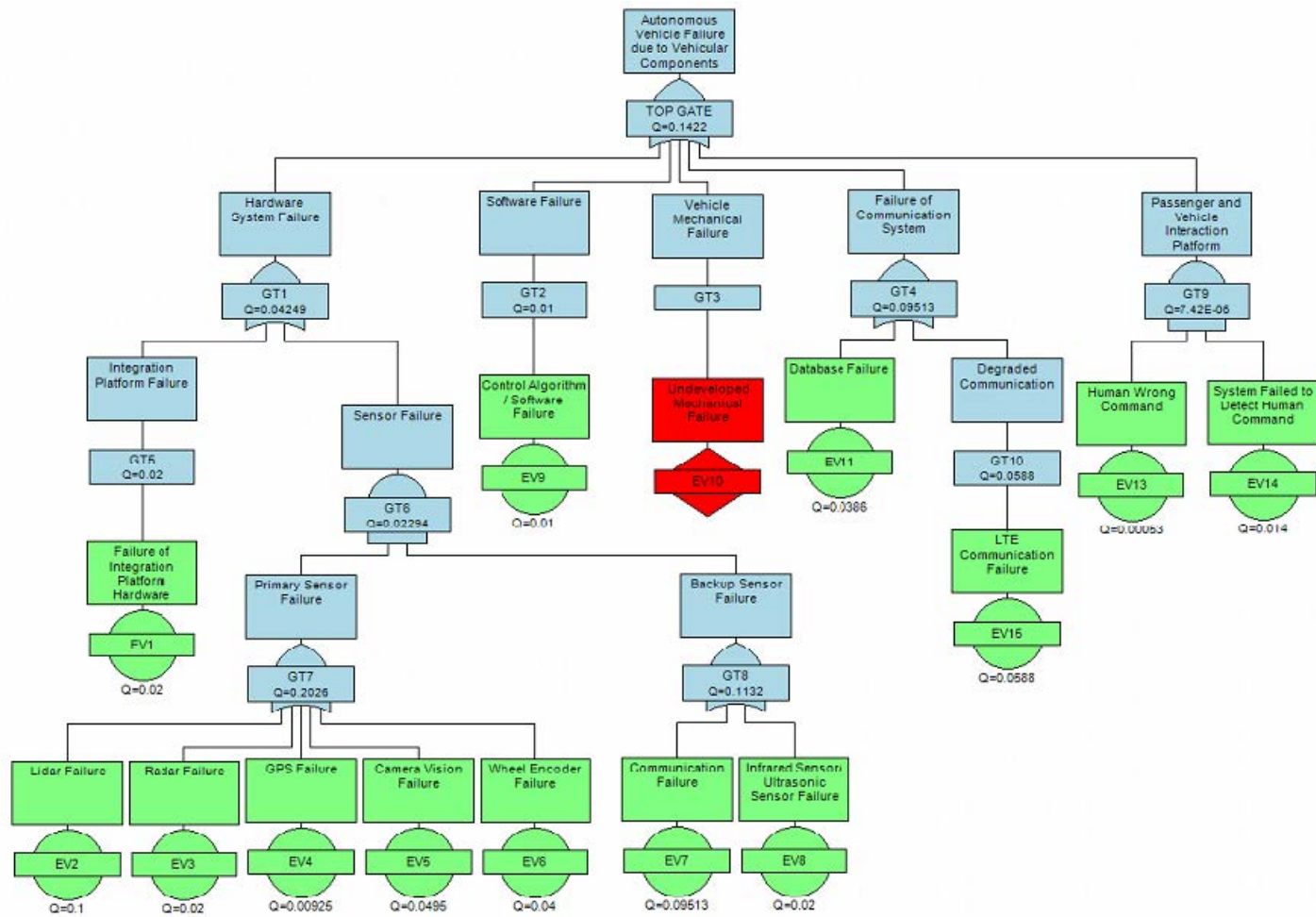


Figure 2.2: FTA considering failure due to vehicular components. Reproduced from UTRC [65].

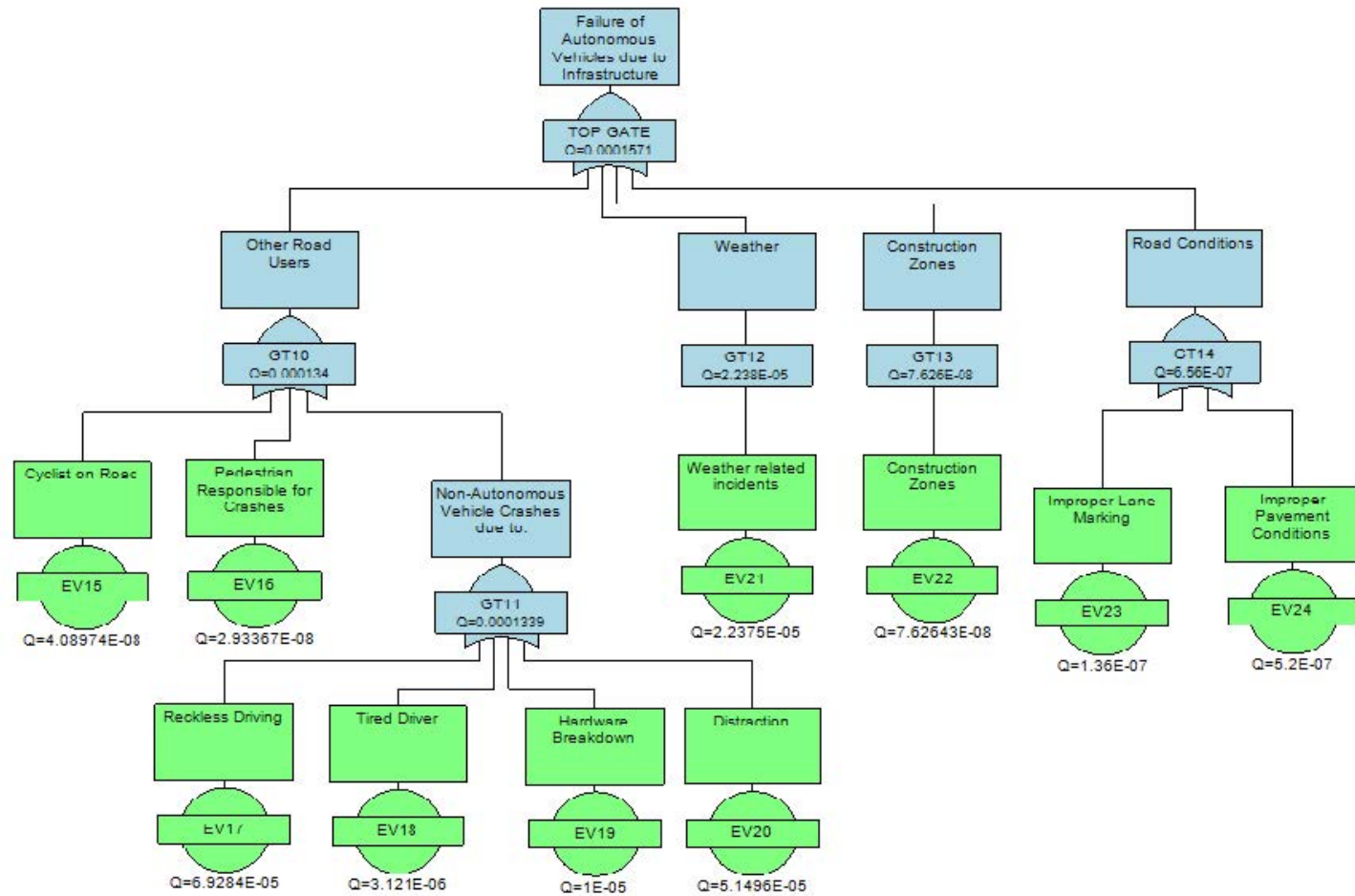


Figure 2.3: FTA considering failure due to infrastructure. Reproduced from UTRC [65].

2.3.3 Systems Theoretic Process Analysis

FMEA and FTA are traditional methods of reliability engineering designed for the systems of the mid-20th century, but some argue that these methods are insufficient for the much more dynamic and complex systems of today. In 2011, Dr. Nancy G. Leveson published a new accident causality model called Systems-Theoretic Accident Model and Processes (STAMP) in her book *Engineering a Safer World: Systems Thinking Applied to Safety*. In contrast to the classical view of accident causality, in which accidents are caused by a directly related series of events, STAMP posits that accidents are the result of complex dynamic processes. STAMP stresses that unsafe conditions can arise even when the individual components of a system are highly reliable, and that accidents often result from an error in the controller's process model, or internal estimation of the state of the system. The systems-based approach to safety better accommodates modern dynamic systems, and also incorporates societal safety structures into the system, as shown in the general model in Figure 2.4. STAMP provides the foundation for a number of STAMP-based processes, including a hazard analysis procedure, Systems Theoretic Process Analysis (STPA) [39].

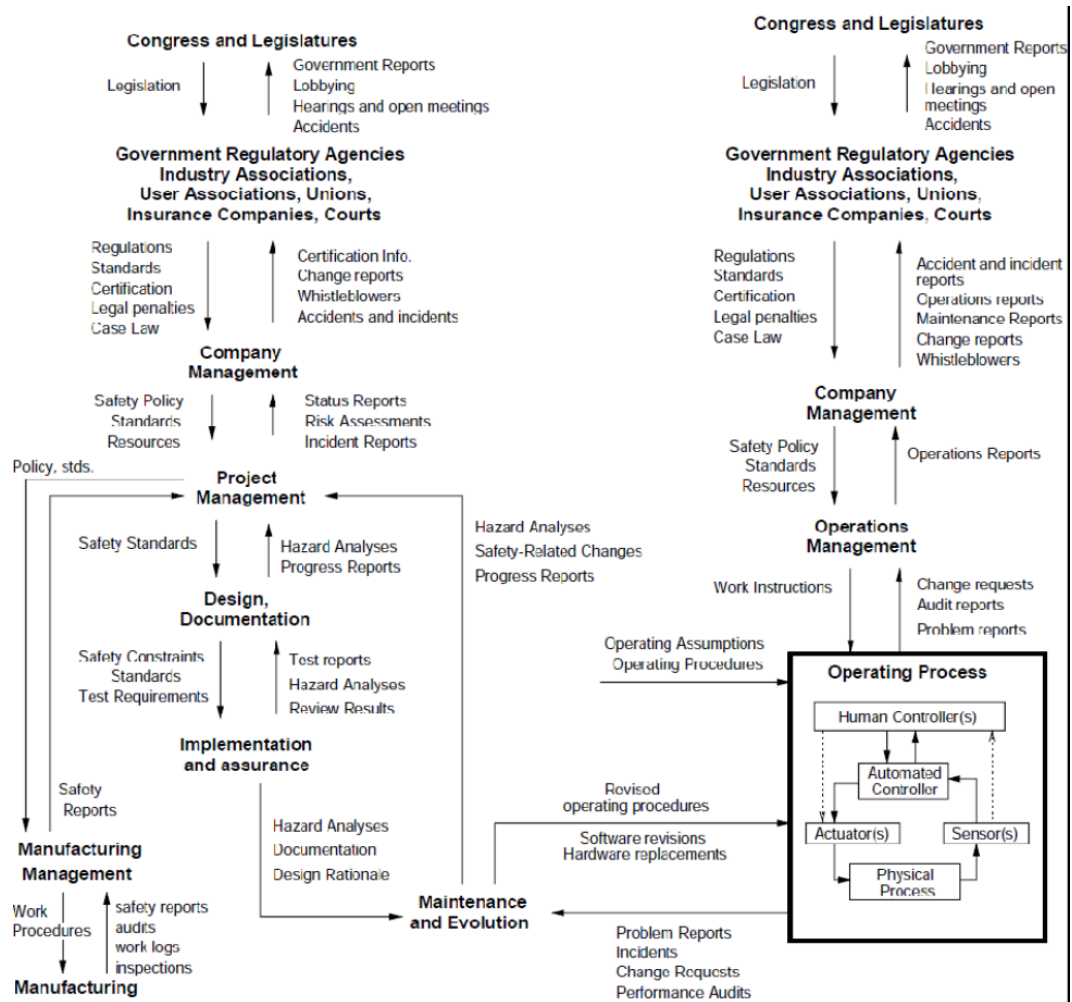


Figure 2.4: General model of a STAMP system. Reproduced from Leveson [39].

In contrast to the linear causality of traditional hazard analysis, the focus of STPA is on identifying control actions that lead to hazards, and correcting flaws in the underlying controller of the system to remove or mitigate them. STPA is performed using two main steps:

Step 1: Identify unsafe control actions.

Step 2: Identify causal factors and control flaws.

STPA has been used academically to analyze AV systems, but has not yet achieved widespread industry use. The following section shows an example of STPA applied to an automated parking assist maneuver (APA) performed by a Level 3 AV [22].

In order to identify unsafe control actions, the analyst must first identify possible accidents, hazards, and safety constraints. In the STAMP nomenclature, accidents are the potential losses from unintended events, hazards are the unsafe conditions that give rise to accidents, and safety constraints are restrictions that prevent hazards. The accidents, hazards, and safety constraints for the APA maneuver are shown in Tables 2.3 and 2.4.

Table 2.3: Accidents for the example APA system. Reproduced from France [22].

A-1	Death, injury, or property damage resulting from a collision with a person, vehicle, object, or terrain.
A-2	Injury or property damage occurring within the vehicle, without a collision.
A-3	Loss of customer satisfaction with automated parking, without injury or property damage.

Table 2.4: Hazards and safety constraints for the example APA system. Reproduced from France [22].

System-Level Hazards		System Safety Constraints	
H-1	The vehicle does not maintain a safe minimum distance between itself and obstacles such as pedestrians, vehicles, objects, and terrain. [A-1].	SC-1	The vehicle must maintain a safe minimum distance between itself and obstacles such as pedestrians, vehicles, objects, and terrain.
H-2	Occupants or cargo are subjected to sudden high forces that may result in injury or property damage. [A-2]	SC-2	The vehicle must not brake, accelerate, or turn at speeds that would result in injury or property damage.
H-3	The vehicle parks inappropriately, either in an unsuitable space (e.g. blocking a fire hydrant) or in violation of parking guidelines (e.g. excessively far from the curb). [A-3]	SC-3	The vehicle must park in valid, legal spaces and at an appropriate distance to the curb.

Next, the control structure of the system is developed. The control structure identifies the components of the system and indicates the signals that are passed between them. The control structure of the APA system is shown in Figure 2.5.

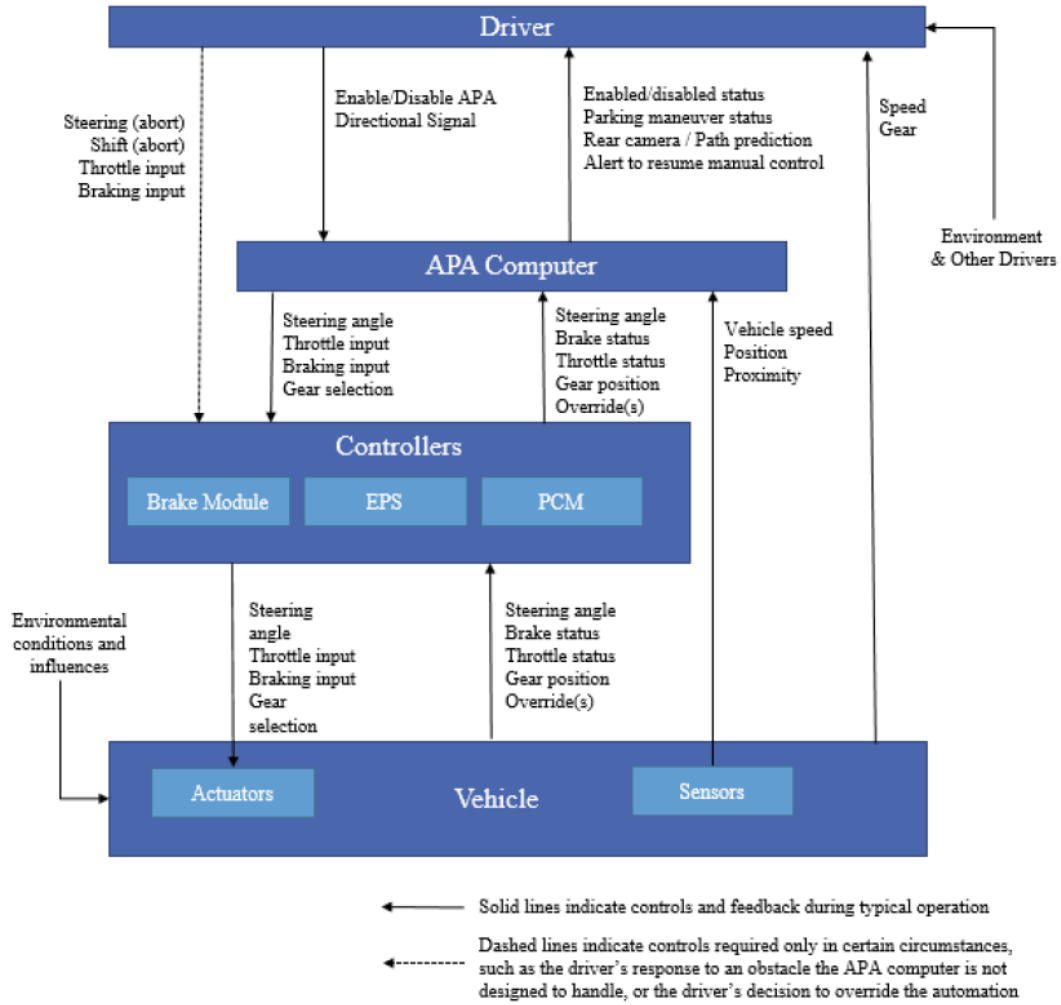


Figure 2.5: Control structure for the example APA system. Reproduced from France [22].

Step 1 is completed by identifying control actions in the control structure that can lead to hazards. These actions are called unsafe control actions (UCAs), and they occur when the safety constraints are insufficient or unenforced. UCAs come in four general forms [39]:

1. A control action required for safety is *not* provided or not followed.
2. An unsafe control action is provided.
3. A potentially safe control action is provided too early or too late, that is, at the wrong time or in the wrong sequence.
4. A control action required for safety is stopped too soon or applied too long.

An excerpt of the UCAs for the APA maneuver related to braking are shown in Table 2.5.

The second and final step is to examine each UCA, determine what flaw in the control structure could cause it to occur, and introduce an additional control or safety measure to enforce safety constraints.

Table 2.5: Excerpt of Unsafe Control Actions (UCAs) for the example APA system. Reproduced from France [22].

	Not Providing Causes Hazard	Providing Causes Hazard	Incorrect Tim- ing/Order	Stopped Too Soon / Applied Too Long
Brake (APA computer)	UCA 3-25: APA computer does not brake when braking is necessary to prevent collision. [H-1]	UCA 3-26: APA computer brakes when APA is disabled. [H-1] UCA 3-27: APA computer brakes when doing so creates an obstruction. [H-1] UCA 3-28: APA computer brakes when doing so exposes the occupants and cargo to sudden high forces. [H-2]	UCA 3-29: APA computer brakes too soon to complete the maneuver. [H-3] UCA 3-30: APA computer waits too long to brake to avoid collision. [H-1]	UCA 3-31: APA computer continues braking for too long and stops short of completing the maneuver. [H-3] UCA 3-32: APA computer does not brake for long enough to avoid collision or stop within desired bounds. [H-1]

2.4 Summary

Real-world testing, simulation, FMEA, FTA, and STPA are five methods that have been used to assess the safety of AV systems. Real-world testing is capable of rigorously proving safety, but requires a data set that could take decades to acquire. Simulation appears to eliminate this challenge by testing the vehicle’s controller in a safe, virtual environment. FMEA is a qualitative, bottom-up approach to identifying and organizing failure modes; in contrast, FTA provides a quantitative, top-down approach that represents interdependencies between failure modes. STPA identifies unsafe control actions using the systems-based STAMP model, which may be better suited to highly dynamic systems such as AVs in traffic. The advantages and disadvantages of each method will be further explored in the following chapters.

While this chapter presents analyses performed by various researchers on various AV systems, the remainder of this thesis will demonstrate how these methods can be performed on a specific system, so that they can be more effectively compared. The following chapter details the approach taken to implement a simulation and three analytical methods on the case of a fully autonomous vehicle performing an unprotected left turn at an intersection.

Chapter 3

Methodology

In this thesis, each method described in the previous chapter, with the exception of real-world testing, will be applied to the specific maneuver of an AV performing an unprotected left turn at an intersection with heavy traffic. This chapter will introduce the details of the simulation, and describe the approach made to the analytical methods.

3.1 Simulation

A simulation, adapted and expanded from MATLAB code supplied in class materials, was built to model an AV making an unprotected left turn at an intersection with heavy traffic. It should be noted that the goal of this simulation is not to assess the safety of a particular AV design, nor to accurately and comprehensively model a real-world driving environment. The objective of this simulation is to explore and assess the ability of simulation tools to verify safety in AVs, evaluate the strengths and weaknesses of simulation in comparison to and in the context of analytical methods such as FMEA, FTA, and STPA, and provide insight into what method or methods might be most useful in AV design and risk assessment. This section will establish the parameters of that simulation including the dynamic vehicle model, the test

conditions, the controller, and the outputs of interest.

3.1.1 Vehicle Model

The model used in this simulation is governed by the following dynamic equations. The model has four states: the position in 2D space, x and y , the angle between the longitudinal axis of the vehicle and the global x -axis, ϕ , and the velocity in the vehicle's forward (longitudinal) direction, v . The model has two control inputs: the force applied at the center of mass along the longitudinal axis, $u_1 = F$, and the angle between the front wheel and the longitudinal axis, $u_2 = \delta$.

$$\dot{x} = v \cos \phi$$

$$\dot{y} = v \sin \phi$$

$$\dot{\phi} = \frac{v}{L} \tan u_2$$

$$\dot{v} = u_1/m$$

This model includes several simplifications. First, steering is represented by a single wheel. Second, all longitudinal traction forces are represented as a single force acting at the center of mass and in the direction of travel. The result is that the wheels are assumed not to slip in either the longitudinal or lateral directions, so that the vehicle has kinematic steering and dynamic acceleration. This ignores some possible failure modes such as dynamic instability due to insufficient cornering forces. These simplifications

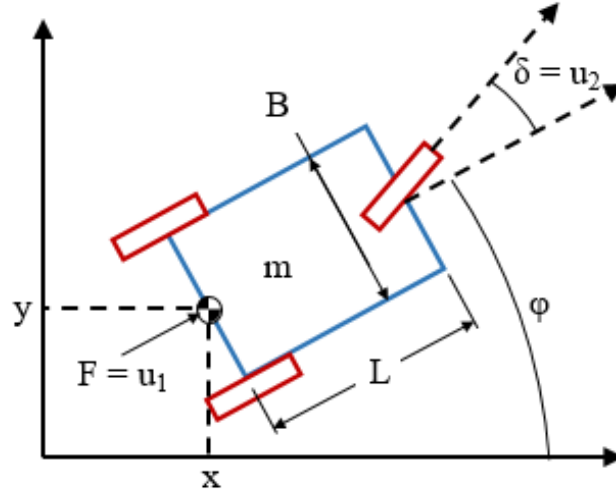


Figure 3.1: Variables of interest in the vehicle model used in this simulation.

Table 3.1: Specifications for the simulated AV, based on the Chrysler Pacifica used by Waymo [1].

Gross Weight	m	6300 lbs (2857 kg)
Wheelbase	L	121.6 in (3.089 m)
Width	B	79.6 in (2.02 m)

are acceptable because the AV performs the turn at low speed, where dynamic instability is unlikely to occur.

The parameters of the vehicle are based on the Chrysler Pacifica minivan used by Waymo in its Waymo One AV taxi service in Phoenix, Arizona. Relevant specifications of that vehicle are depicted in Table 3.1.

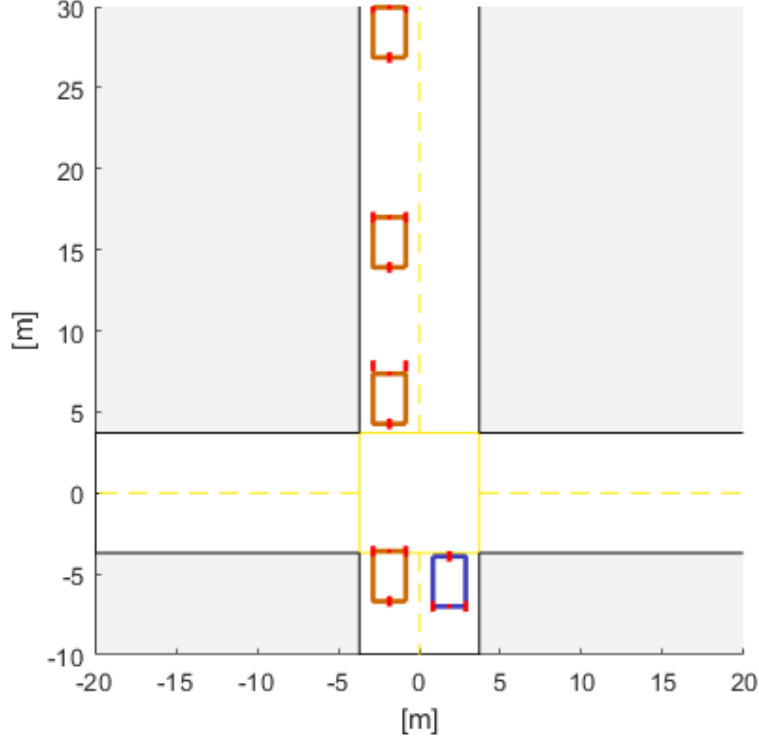


Figure 3.2: Layout of the simulated intersection, with AV (blue) waiting to begin left turn across oncoming cars (orange).

3.1.2 Test Conditions

Although in reality, AVs are required to navigate a wide variety of complex situations, this simulation focuses specifically on the unprotected left turn at a busy intersection. This maneuver was chosen because it is often cited as a particularly difficult task for AVs [18].

The simulated area consists of an intersection between a pair of two-lane roads. The lanes are 3.7 m (12 ft) wide in accordance with standard U.S.

lane widths [9], and the cars travel at a desired speed of 10 m/s (22 mph). The flow of traffic is randomly generated, with each car having a following time of 0.5s to 3s (evenly distributed) behind the car in front of it.

The oncoming cars are equipped with the PD controller shown below to regulate their speed. The subscript a refers to the current car, and the subscript b refers to the car in front. The controller is designed to converge to a following distance of 2 seconds of travel plus 2 times the length of a vehicle, such that vehicles will adjust their speed to gradually close gaps and reduce tailgating. The oncoming cars will not steer outside of their lane. The controller outputs for forward acceleration $u_1 = F$ and steering angle $u_2 = \delta$ are given as follows:

$$u_1 = F = 5 * (y_a - y_b - 2 * v_b - 2 * L) + 30 * (v_a - v_b)$$

$$u_2 = \delta = 0$$

3.1.3 AV Controller

The AV begins at rest, outside of the intersection. The controller has a single control action, which is to perform an open-loop left turn into the correct lane. The phases of the turn are shown in Table 3.2.

It is known that the AV crosses the center of the oncoming lane 3.6 seconds after initiating the left turn maneuver. To determine when to initiate the left turn, the controller calculates the expected position of each oncoming

Table 3.2: Open-loop turn sequence executed by the AV.

	$u_1 = F$	$u_2 = \delta$
$t < 2.07s$	4500 N	0
$2.07s < t < 4.25s$	1500 N	30°

car 3.6 seconds from the current time, based on each car’s current speed. The maneuver begins if no expected position falls within a band extending 15m from the possible collision point in either direction.

3.1.3.1 Failure Modes

With accurate information about the environment, the simple controller completes the maneuver successfully with a 100% success rate. To model sensor error, the AV controller only has access to position and velocity data about the oncoming cars that has been adjusted by some random value at the beginning of the simulation. Additionally, the sensors have a chance of completely failing to recognize an oncoming vehicle, to represent failures occurring due to poor weather, vehicles obscured by obstructions, and other total failures.

Failure modes can also occur in the controller itself. In reality, AV controllers are implemented as complex machine learning algorithms, which are difficult to model accurately in this simplified simulation. An attempt to design and implement a realistic AV controller in this study would yield failure modes due to the designer’s inadequate controller design, rather than failure modes representative of real controllers. Instead, the simple controller with a zero-percent failure rate described above is used, and failure modes following

Table 3.3: Failure modes applied to the simulated AV controller, in the general form outlined in STPA [39].

Generalized Failure	AV Failure
A control action required for safety is <i>not</i> provided or not followed.	Type I Controller Error: The vehicle does not initiate a turn at the first available window, resulting in unnecessary delay.
An unsafe control action is provided.	Type II Controller Error: The vehicle initiates a turn immediately, regardless of whether an appropriate window exists.
A potentially safe control action is provided too early or too late, that is, at the wrong time or in the wrong sequence.	Type III Controller Error: The window checked by the AV is shifted by a random distance, such that the turn is initiated too early or too late.
A control action required for safety is stopped too soon or applied too long.	No equivalent.

the general form of unsafe control actions in STPA [39] are applied a known percentage of the time, as shown in Table 3.3. Note that because the control action of initiating a term is applied as an impulse, it cannot be stopped too soon or applied too long, so there is no equivalent failure mode.

3.1.3.2 Statistics

For each test case, the simulation was run 1,000 times to estimate the collision rate. The margin of error e for a proportion estimate of an infinite population with sample size n , population proportion p , and Z-score $z_{\alpha/2}$ is estimated as follows [36].

$$e = \sqrt{\frac{p \cdot (1 - p) \cdot z_{\alpha/2}^2}{n}}$$

In this case, the population proportion p is the percent of all trials under given conditions that end in a collision. This collision rate is estimated using $n = 1,000$ simulated trials. For a 95% confidence interval, $z_{\alpha/2} = 1.96$. The proportion of collisions ending in collision is unknown, so $p = 0.5$ is used conservatively. Therefore, the margin of error in the estimated collision rate is $\pm 3.1\%$.

3.2 Analytical Methods

In addition to a computer simulation, the analytical methods discussed in Chapter 2 were also applied to the left turn scenario. Comparison of the analyses to the simulation will be discussed in Chapter 5. Each analysis was performed on a hypothetical fully-autonomous (SAE Level 5) AV which uses a combination of visual cameras, LIDAR, RADAR, and GPS to estimate its position and surroundings.

3.3 Summary

This chapter describes the approach used to implement four methods on the specific case of a fully autonomous AV making an unprotected left turn at a high-traffic intersection. Technical details for the simulation are presented, including the dynamic vehicle model, the layout of the intersection, and the controllers for the AV and the oncoming cars. Results of these approaches are presented in the following chapter, with discussion presented in Chapter 5.

Chapter 4

Results

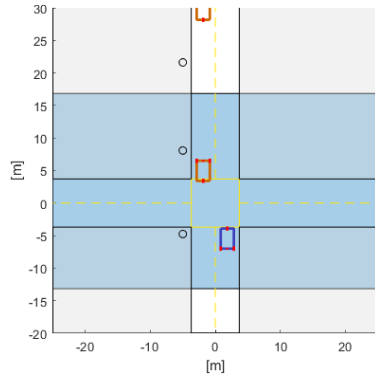
4.1 Simulation

The simulation was run for a variety of test conditions to demonstrate how simulation technology might be used in industry and explore the limitations of the approach. A successful sample trial of the left turn maneuver is shown in Figure 4.1, in contrast to a sample of a collision caused by failure of the sensors to detect an oncoming vehicle shown in Figure 4.2.

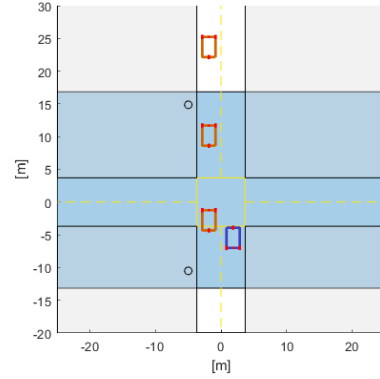
Table 4.1 shows the system failure rate (i.e. collision rate) after 1,000 trials for test cases with varying parameters for each failure modes. Recall from Table 3.3 that errors in the controller occur in three forms: Type I, in which the controller does not turn at the first appropriate window; Type II, in which the controller turns when an appropriate window does not exist; and Type III, in which the AV checks a window offset by a random distance, such that the turn occurs with poor timing. Parameters are given as the percentage chance of the error occurring, or the maximum error in a uniform distribution centered on zero, as indicated.

Figure 4.3 plots the results of 36 test cases that show the interaction between varying maximum velocity error and maximum position error.

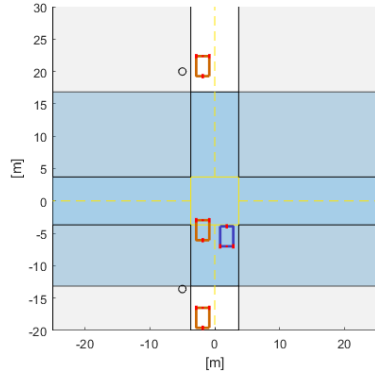
Figure 4.1: Sample of a successful trial of the simulation.



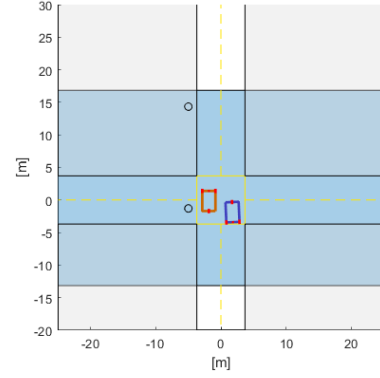
(a) $t = 0s$. The AV (blue) waits to turn across oncoming traffic (orange). Expected vehicle positions in 3.6s shown in black circles, offset.



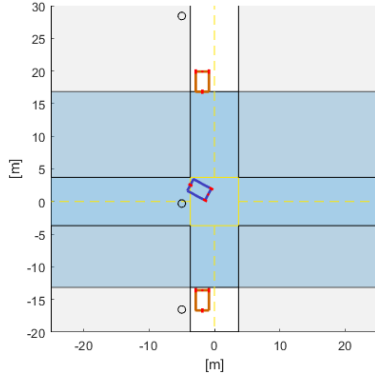
(b) $t = 3.3s$. Turn will not occur if any expected position of any oncoming car in 3.6s falls within the 30m blue shaded region.



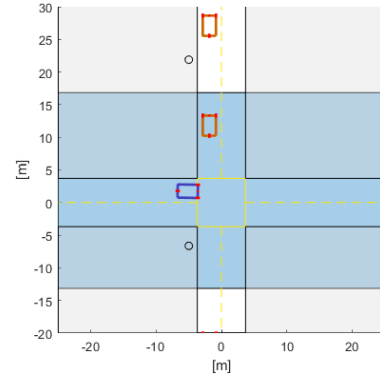
(c) $t = 6.15s$. No oncoming vehicle is expected to be within the shaded region when the AV crosses the lane, so a turn is initiated.



(d) $t = 8.25s$. After accelerating into the intersection, the AV begins to steer left.

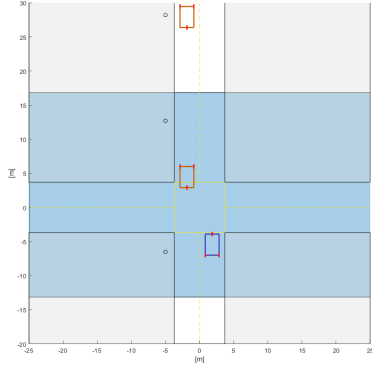


(e) $t = 9.75s$. The AV crosses the oncoming lane. The oncoming cars are at the expected positions shown in (c).

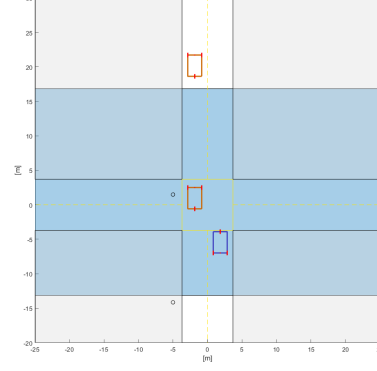


(f) $t = 10.4s$. The AV successfully completes the turn.

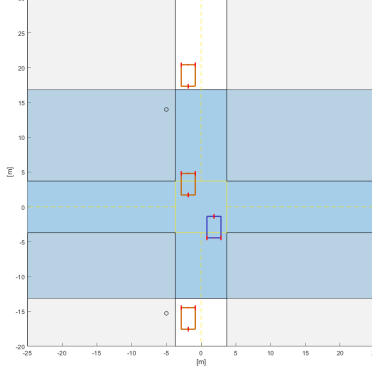
Figure 4.2: Sample of trial of the simulation ending in collision caused by the sensor failing to detect an oncoming vehicle.



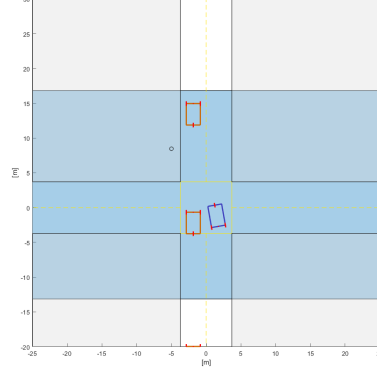
(a) $t = 0s$. The AV (blue) waits to turn across oncoming traffic (orange). Expected vehicle positions in 3.6s shown in black circles, offset.



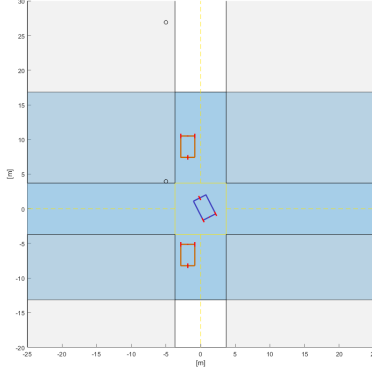
(b) $t = 2.8s$. Turn is initiated despite an expected position falling within the shaded region, due to sensor detection error.



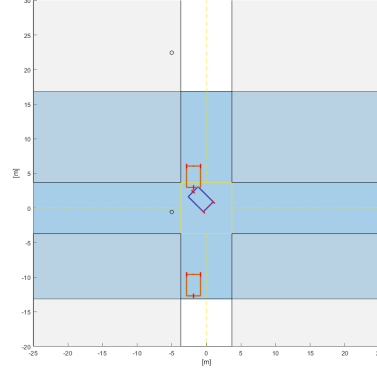
(c) $t = 4.5s$. Vehicle enters intersection.



(d) $t = 5.0s$.



(e) $t = 5.5s$.



(f) $t = 5.9s$. Collision occurs.

Table 4.1: Collision rate and time taken to turn for different combinations of failure mode parameters.

Test Case	Max Position Error [m]	Max Velocity Error [m/s]	Sensor Read Failure [%]	Controller Error, Type I [%]	Controller Error, Type II [%]	Controller Error, Type III [m]	Collision Rate [%]	Time to turn [s]
Baseline								
1	0	0	0	0	0	0	0	12.8±14.2
Isolated failure modes								
2	5	0	0	0	0	0	0	10.2±10.8
3	7	0	0	0	0	0	18.4	8.9±9.2
4	10	0	0	0	0	0	47.3	6.9±6.8
5	0	1	0	0	0	0	0	11.2±12.2
6	0	2	0	0	0	0	25.9	8.6±8.9
7	0	3	0	0	0	0	52.6	5.6±7.2
8	0	0	100	0	0	0	81	0.05±0
9	0	0	0	100	0	0	0	24.4±18.8
10	0	0	0	0	100	0	80.5	0.01±0
11	0	0	0	0	0	5	0	12.2±12.7
12	0	0	0	0	0	10	31.3	12.4±12.7
13	0	0	0	0	0	15	51.9	12.6±13.2
Combined failure modes								
14	5	1	10	0	0	0	31.7	6.5±7.1
15	0	0	0	10	10	0	37.4	12.1±14.1
16	0	0	10	100	0	0	22.3	19.5±12.6
17	2	1	0	0	0	0	0	11.1±11.0
18	2	1	0	0	0	1	22.1	7.3±8.0
19	1	0.5	10	10	10	1	28.4	7.1±8.8
20	0.5	0.25	1	1	1	0.5	4.0	11.2±12.1

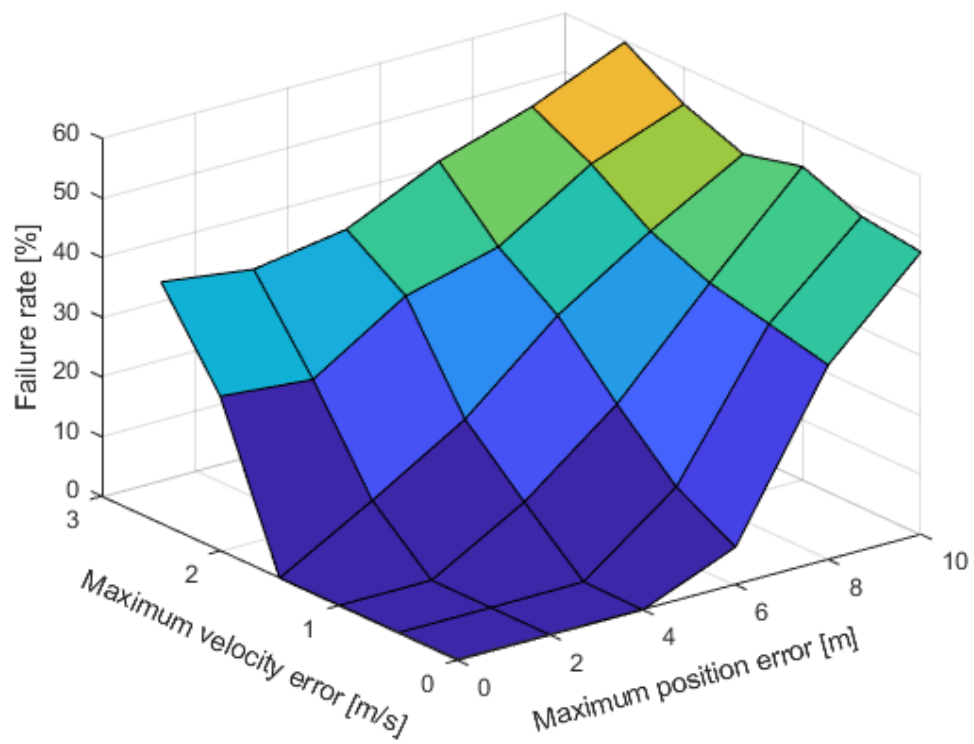


Figure 4.3: Interacting effects of sensor error in position and velocity measurements.

4.2 Analytical Methods

In addition to a computer simulation, the analytical methods discussed in Chapter 2 were also applied to the left turn scenario. The results are shown below. Comparison of the analyses to the simulation will be discussed in the following chapter.

4.2.1 Failure Modes and Effects Analysis

FMEA was performed for the left turn maneuver, and the results are shown in Figures 4.4 and 4.5. Note that the rating scale for severity, occurrence, and detectability can differ between systems, since different systems have different stakes. The rating scale used in this FMEA was designed for the AV left turn maneuver and is shown in Table 4.2.

Table 4.2: Ratings used for the FMEA analysis.

	Severity (SEV)	Occurrence (OCC)	Detectability (DET)
1	No significant effect. Customer does not notice failure.	Failure rate is very low and supported by data.	Condition will always be detected before failure occurs.
2	Customer notices failure, possibly experiences annoyance.	Failure rate is low and supported by data.	Condition will nearly always be detected before failure occurs.
3	Loss of customer satisfaction.	Failure rate is expected to be low, but not supported by data.	Condition is very likely to be detected before failure occurs.
4	Property damage is possible if other redundant components also fail.	Failure rate is moderate and supported by data.	Condition is likely to be detected before failure occurs. Automatic detection methods can be used.
5	Property damage is possible without other failures.	Failure rate is expected to be moderate, but not supported by data.	Condition is likely to be detected before failure occurs. Must be detected by manual inspection.
6	Property damage is likely.	Failure is likely to occur over life cycle of vehicle.	Condition more likely than not to be detected before failure occurs.
7	Collision endangering humans is possible if other redundant components also fail. Property damage is nearly certain.	Condition is expected to occur once in a matter of years.	Condition may be detected before failure occurs (approximately 50% chance).
8	Collision endangering humans is possible without other failures. Property damage is nearly certain.	Condition is expected to occur once in a matter of months.	Condition may be detected before failure occurs (approximately 50% chance).
9	Collision endangering humans is likely.	Condition is expected to occur once in a matter of weeks.	Condition unlikely to be detected.
10	Collision endangering humans is nearly certain.	Condition occurs very often (e.g. daily) and regularly.	Condition cannot be detected by current means.

Component	Failure Mode	Failure Effects	Sev	Potential Causes	Occ	Det	Recommended Action	RPN
Sensors								
Vision-based camera	Poor visibility		5	Driving at night, poor weather (heavy rain, snow, or fog), dirt or obstruction over lens	10	2	If confidence in sensor data is low, pull over or alert human driver to take control	100
	Hardware failure		5	Manufacturing fault, or at end of life cycle	4	4	Annual inspection	80
LIDAR	Poor visibility		5	Poor weather (heavy rain, snow, or fog), dirt or obstruction over sensor	8	2	If confidence in sensor data is low, pull over or alert human driver to take control	80
	LIDAR interference		5	Other AVs in the area using LIDAR	10	2	Laser signal should be coded with ID to prevent interference	100
	Positional error (bias error or noise)		4	Intrinsic to sensor	10	2	Measurement uncertainty should be conveyed to decision-making algorithm	80
	Hardware failure		5	Manufacturing fault, or at end of life cycle	3	4	Annual inspection	60
RADAR	RADAR interference	Outcome depends on whether other sensors remain operational and how the controller compensates for the loss of data. Collision is possible.	5	Other AVs in the area using RADAR	10	2	RADAR signal should be coded with ID to prevent interference	100
	Positional error (bias error or noise)		4	Intrinsic to sensor	10	2	Measurement uncertainty should be conveyed to decision-making algorithm	80
	Hardware failure		5	Manufacturing fault, or at end of life cycle	3	4	Annual inspection	60

Figure 4.4: FMEA analysis for the vehicle completing a left turn.

Component	Failure Mode	Failure Effects	Sev	Potential Causes	Occ	Det	Recommended Action	RPN
Controller								
Object Identification	Object misidentified	Varies from mild to severe, depending on misidentification. Collision possible.	8	Camera/LIDAR inoperable and RADAR resolution too low for object ID, error by machine learning algorithm	4	10	If confidence in sensor data is low, pull over or alert human driver to take control. Ensure machine learning training is adequate.	320
Path Planning	Planned path exits road	Varies from mild to severe, depending on surroundings. Collision possible.	8	Error by machine learning algorithm	2	10	Ensure machine learning training is adequate	160
	Cyberattack	Varies depending on intent of attacker. Potential for collision, kidnapping, or other serious crimes.	10	Malicious virus, cybersecurity flaw	5	4	Cybersecurity software with frequent updates	200
Decision Making	Turn initiated when appropriate window does not exist	Collision nearly certain	10	Error by machine learning algorithm	2	10	Ensure machine learning training is adequate	200
	Turn not initiated when appropriate window exists	Delay, loss of customer satisfaction	3		10	10		300
	Turn initiated with appropriate window, but poor timing	Collision likely	9		4	10		360
Vehicle failure								
Tires	Tire blowout	Depends on surroundings and location of failed tire. Collision possible.	8	Tire at end of life cycle, manufacturer fault, sharp objects in road, overloaded vehicle, or low tire pressure	7	3	Annual inspection	168
Brakes	Brake failure	Depends on surroundings. Collision likely.	9	Leak in brake fluid line or loss of air pressure in brake line	6	3	Annual inspection	162
Power	Power loss	Vehicle ceases all operation. Collision possible. Vehicle is stranded on road leading to delay and loss of customer satisfaction.	9	User failure to refuel or recharge vehicle.	6	2	Alert user when power is low. Before complete power loss, autonomously pull over to safe location.	108

Figure 4.5: FMEA analysis for the vehicle completing a left turn, continued.

4.2.2 Fault Tree Analysis

FTA was performed for the left turn maneuver, the results of which are shown in Table 4.3 and Figure 4.6. Estimated values are used for basic event probabilities.

Table 4.3: Probability of basic, intermediate, and top-level failure events during the left turn maneuver. Probabilities for basic events are estimated, probabilities for intermediate and top-level events are calculated.

Event	Type	Probability
Collision	Top-Level	1.35e−5
Oncoming car fails to react	Basic	0.5
AV failure	Intermediate	2.71e−5
Vehicle mechanical failure	Intermediate	6.00e−8
Tire blowout	Basic	3.00e−8
Brake failure	Basic	2.00e−8
Other mech. failures	Basic	1.00e−8
Sensor failure	Intermediate	1.50e−5
Significant bias error	Basic	2.00e−8
Failure to detect obstacle	Intermediate	1.50e−5
LIDAR failure	Intermediate	1.00e−2
Inclement weather	Basic	0.01
LIDAR hardware failure	Basic	3.00e−8
RADAR failure	Intermediate	3.00e−2
RADAR alone unable to ID obstacle	Basic	0.03
RADAR hardware failure	Basic	3.00e−8
Camera failure	Intermediate	5.00e−2
Poor visibility	Basic	0.05
Camera hardware failure	Basic	3.00e−8
Controller failure	Intermediate	1.20e−5
Object identification	Basic	3.00e−6
Path planning	Basic	3.00e−6
Decision making	Intermediate	6.00e−6
Turn occurs regardless of surroundings	Basic	3.00e−6
Turn occurs with incorrect timing	Basic	3.00e−6

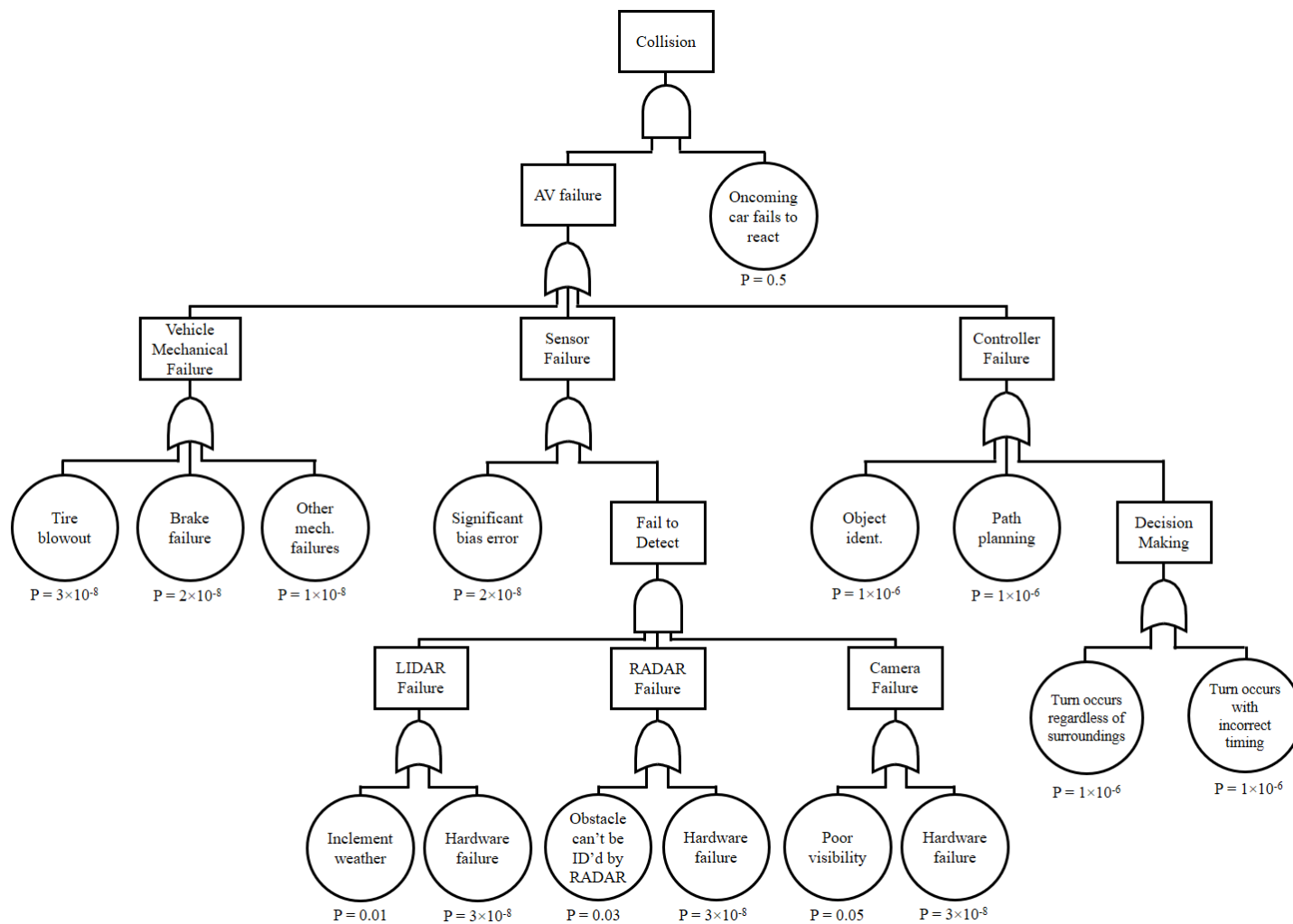


Figure 4.6: Fault tree analysis performed for the left turn maneuver.

4.2.3 Systems Theoretic Process Analysis

STPA was performed for the left turn maneuver. First, accidents are defined as shown in Table 4.4 and hazards and safety constraints are defined as shown in Table 4.5. Note that some accidents, hazards, and safety constraints are the same as or adapted from those listed developed by France [22], as reproduced in Chapter 2 of this thesis.

Table 4.4: Accidents for the left turn maneuver. Adapted from France [22].

A-1	Death, injury, or property damage resulting from a collision with a person, vehicle, object, or terrain.
A-2	Injury or property damage occurring within the vehicle, without a collision.
A-3	Loss of customer satisfaction with the vehicle, without injury or property damage.

Table 4.5: Hazards and safety constraints for the left turn maneuver. Adapted from France [22].

System-Level Hazards		System Safety Constraints	
H-1	The AV initiates a turn timed such that its trajectory intersects with the trajectory of an oncoming vehicle. [A-1].	SC-1	The AV must ensure an appropriate window between oncoming cars exists at the time that it traverses the oncoming lane.
H-2	Occupants or cargo are subjected to sudden high forces that may result in injury or property damage. [A-2]	SC-2	The vehicle must not brake, accelerate, or turn at speeds that would result in injury or property damage.
H-3	The AV takes too long to initiate a turn, causing unnecessary delays. [A-3]	SC-3	The AV must have a low rate of allowing acceptable windows between oncoming vehicles to pass without initiating the turn.
H-4	The AV violates traffic laws, resulting in fines or general confusion. [A-3]	SC-4	The AV must observe and obey all local traffic laws.

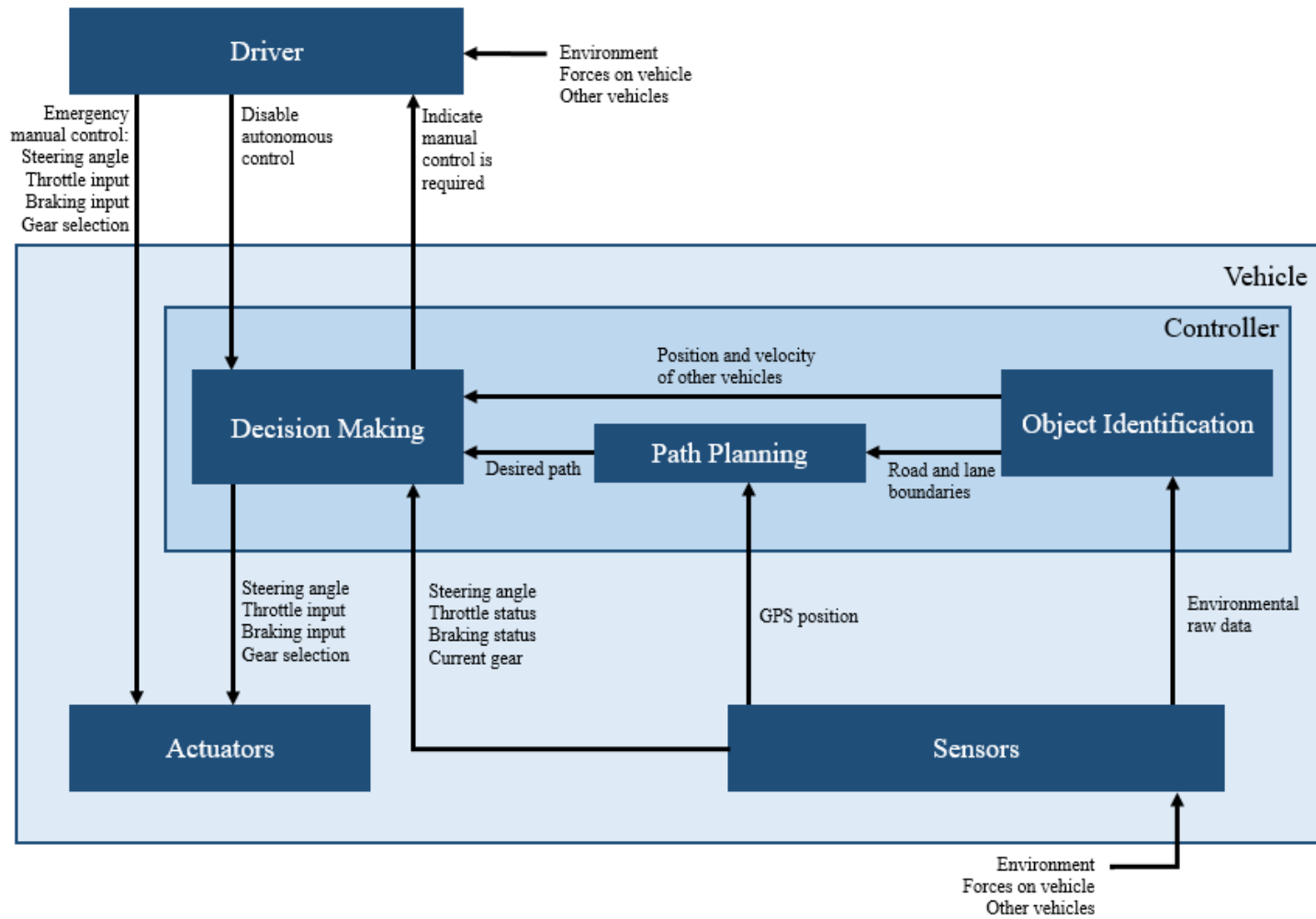


Figure 4.7: Control structure for the AV system. Adapted from France [22].

Table 4.6: Unsafe Control Actions (UCAs) for the left turn maneuver.

Control Action	Not Providing Causes Hazard	Providing Causes Hazard	Incorrect Timing/Order	Stopped Too Soon/Applied Too Long
Initiate Left Turn	UCA-1: AV does not initiate turn when an appropriate window exists. [H-3]	UCA-2: AV initiates turn when an appropriate window does not exist, putting vehicle on collision path. [H-1]	UCA-3: AV initiates turn too early when a window exists, putting vehicle on collision path with front car. [H-1] UCA-4: AV initiates turn too late when a window exists, putting vehicle on collision path rear car. [H-1]	
Brake	UCA-5: AV does not brake at start of turn, such that the turn is taken too quickly for passenger comfort. [H-2]	UCA-6: AV brakes unexpectedly such that the vehicle stops on a collision path with oncoming cars. [H-1] UCA-7: AV brakes unexpectedly, uncomfortably jarring passengers. [H-2]	UCA-8: AV brakes while accelerating, causing unnecessary wear and damage to brakes. [H-2]	
Continued on next page				

Table 4.6 (continued)

Accelerate	UCA-9: AV does not enter intersection at start of turn, missing an opportunity. [H-3] UCA-10: AV does not accelerate out of turn, traveling well below desired speed. [H-3]	UCA-11: AV enters intersection without a signal to initiate turn, violating traffic laws. [H-4]		UCA-12: AV is traveling too slow as it enters the turn, possibly resulting in collision. [H-1] UCA-13: AV is traveling too quickly as it enters the turn, resulting in discomfort. [H-2] UCA-14: AV accelerates too much after the turn, exceeding the road's speed limit. [H-4]
Steer left	UCA-15: AV fails to turn and instead proceeds straight through the intersection, creating a delay for passengers. [H-3]	UCA-16: AV turns at an inappropriate time, entering a lane of oncoming traffic. [H-1]		UCA-17: AV turns too far or not far enough during the turn, exiting the lane. [H-1]
Steer right	UCA-18: AV fails to straighten direction after turn, exiting the lane. [H-1]	UCA-19: AV turns at an inappropriate time, exiting the lane. [H-1]		UCA-20: AV turns too far or not far enough when straightening course after the turn, exiting the lane. [H-1]

The next and final step of STPA is to identify how each UCA could occur by examining the control structure, and design safety measures to prevent them from occurring [39]. In Figure 4.7, all relevant control signals are directed from the decision making algorithm to the actuators, and the throttle, braking, and steering commands are subactions under the control action to initiate the left turn sequence. In this case, all UCAs shown in Table 4.6 are caused either by an error in the decision making algorithm itself or can be traced back through the signals in Figure 4.7 to the path planning stage, object identification stage, or sensor output. As an example, possible causes and possible safety measures are listed for UCA-2 are listed in Table 4.7.

Table 4.7: Possible causes and recommended actions for UCA-2: AV initiates turn when an appropriate window does not exist, putting vehicle on collision path.

Possible cause	Recommended action
Misclassification in decision making machine learning algorithm due to insufficient training data.	Verify by separate analysis that machine learning algorithm has sufficient training data to ensure probability of edge case is within known and acceptable limits. Controller should alert driver if prediction confidence is too low.
Error in object identification machine learning algorithm due to insufficient training data.	Verify by separate analysis that machine learning algorithm has sufficient training data to ensure probability of edge case is within known and acceptable limits. Controller should alert driver if prediction confidence is too low.
Error in path planning machine learning algorithm due to insufficient training data.	Verify by separate analysis that machine learning algorithm has sufficient training data to ensure probability of edge case is within known and acceptable limits. Controller should alert driver if prediction confidence is too low.
Inaccurate or insufficient sensor data sent to object identification.	Some combination of LIDAR, RADAR, and camera systems sufficient to identify road boundaries and objects must be operational at all times under all conditions. Verify reliability of these systems by separate analysis.
Desired path exits road or lane boundary due to inaccurate or insufficient GPS data sent to path planning.	All generated paths must remain within lane boundaries regardless of signal from GPS.
Inappropriate and hazardous signal is sent to steering, throttle, or braking actuator because controller process model does not accurately reflect the current state of the steering angle, throttle, or brakes.	Install redundant sensors to steering column, throttle, brakes, etc. as necessary. Verify by separate analysis the reliability of these sensors.

Chapter 5

Discussion

5.1 Simulation

Results from the simulation show some of the types of data that can be gathered from simulation. For example, Figure 4.3 shows the failure rate as a function of sensor error in position and velocity measurements, and identifies a safe region within which collision will not occur due to sensor error. Designers can use this information to ensure the accuracy of the sensors used is within the boundaries of this region. Additionally, Table 4.1 shows the failure rate when each failure mode is isolated. This data helps to prioritize failure modes, similar to the RPN in FMEA, and can guide the efforts of designers working to improve safety.

As shown in Figures 4.1 and 4.2, simulation can store all data from each trial, allowing designers to examine, replay, and alter any given result. When an accident occurs in simulation, not only can the exact cause be determined, but the same conditions can be replayed with minor variations, a process referred to as “fuzzing” [31], in order to analyze particularly hazardous scenarios.

Perhaps the greatest advantage of simulation was not captured by this study due to practical limitations: the ability to work directly with the ma-

chine learning algorithms that govern object identification, path planning, and decision making in AVs. Analytical methods are generally limited in representing these complex algorithms as a black box, which cannot be further deconstructed. Of the methods discussed herein, only simulation and real-world testing can examine how the AV's software actually responds to its environment, and simulation far outperforms real-world testing in terms of time and safety.

However, the simulation also demonstrates a critical flaw in this method. In this simulation, the oncoming cars, representing human drivers, are governed by simple PD controllers which regulate the following distance behind the car in front. This doesn't account for the variability in human driving - in reality, humans drive with varying degrees of aggressiveness, reflected in their desired speed, following distance, lane-changing frequencies, and other behaviors. Even more unpredictable is how a human driver might respond to an AV pulling into their path at an intersection. Although the controllers for the other environmental cars in an actual simulation developed by an AV manufacturer would be far more complete than the simplistic one presented here, the problem remains that in any computer model human drivers must be represented by algorithmic controllers - and designing those controllers is one of the core problems of AV design. Therefore, testing in simulation is built on circular reasoning; it is the attempt to prove that an AV controller is sufficiently human-like by testing it in an environment with other controllers that are assumed to be sufficiently human-like.

Some simulations mitigate this problem by simulating an environment based on real data gathered from public roads [57], but the problem remains - as soon as the AV takes any action, it cannot be predicted how human drivers would respond. Additionally, gathering the data for this form of simulation is nearly as expensive as real-world testing, as it requires humans to drive in cars outfitted with full sensor arrays.

It may be possible to approximate the responses of human drivers using data-based stochastic models. However, proving these models behave in human-like ways leads to many of the challenges discussed in this thesis. Irrespective of the model's accuracy, simulation can ultimately only examine how an AV would behave in an environment populated by other vehicles controlled by algorithms. For this reason, simulation may be most useful in a distant future in which AVs supplant all human-driven vehicles, bypassing many challenges of human-AI interactions.

Another disadvantage of simulation is the substantial time and cost required to both develop and run them. In contrast to the analytical methods presented, which can be performed by an individual or small team with limited resources in a short period of time, a thorough and complete simulation built from the ground up is a significant investment. Because the price of NVIDIA's DRIVE Constellation platform is currently unknown, it is not clear to what degree its release will increase access [47].

After a simulation platform is built or bought, there is significant cost in terms of the time required to run it. Combined, even the simplified simulation

described in this study required over 12 hours to gather the data presented in the previous chapter. Although simulations used in industry are carried out using far more powerful computers than what is available to this researcher, the scope of those simulations is also far larger. Regardless of scale, any simulation will have to simultaneously process multiple controllers and collision detection, both of which can be computationally expensive. This thesis greatly magnified the failure rate of components to reduce the time required to run the simulation, but working with realistic numbers drastically increases the sample sizes, and therefore run times, required. Recall from Section 3.1.3.2 that the margin of error e for a proportion estimate of an infinite population is estimated as follows.

$$e = \sqrt{\frac{p \cdot (1 - p) \cdot z_{\alpha/2}^2}{n}}$$

If the actual probability of a collision during the left turn maneuver is $p = 5 \times 10^{-6}$ (figure estimated for the sake of example), then for a 95% confidence interval and a margin of error of $0.1p$, the sample size required is approximately 77 million trials of the left turn maneuver, compared to the 1,000 trials used in this simulation. These sample sizes may not be practical, even in simulation.

Overall, it is clear that simulation has value in assessing the safety of AVs, by addressing the two major disadvantages of real-world testing: time required and risk to the public. However, there are serious concerns about

simulation’s ability to accurately represent driving environments, and especially the human drivers within them. Further development of controllers representing human drivers, extensive use of “fuzzing” methods, and testing in simulation based on real data rather than fully-simulated environments are likely to yield the most meaningful results in assessing the safety of an AV system.

5.2 Analytical Methods

5.2.1 Failure Modes and Effects Analysis

The FMEA performed in this work identified the components of the controller as being the most significant risks of the AV system. The object identification, path planning, and decision making routines all had failure modes with RPNs of at least 200, while all failure modes of other components were below 200. Most of the risk comes from the fact that controller failure is likely to lead to collision (high severity), and it is difficult to identify that a hazardous decision has been made until damages have occurred (low detectability). Although sensor failure can also lead to collision, the severity is mitigated by redundant sensors, and the failure modes are generally caused by predictable and detectable conditions (e.g. poor weather, inherent measurement uncertainty). Similarly, mechanical failures such as tire blowouts and brake failures are generally detectable and preventable via annual inspection.

These RPN values help the designers to prioritize risks, and the methodical approach can help identify failure modes that had not previously

be considered. Additionally, FMEA can be performed relatively quickly at a low cost, with limited data, and thus has significant value as an approach to reducing risk.

Despite this value, FMEA fails to assess the comparative safety of the system in a useful way. In FMEA, RPNs are commonly calculated once after all failure modes have been identified, then a second time once recommended actions have been taken to minimize risks. As RPNs are only the product of subjective values without a consistent underlying model, they are only useful in comparison to other RPNs (e.g. before and after recommended actions are implemented), and cannot be meaningfully compared to other quantities and other systems. If the goal of assessing AV safety is to determine whether AVs are safer than conventional drivers in terms of accident rates, FMEA offers no meaningful comparison. FMEA performed on a human-controlled vehicle would yield a completely different set of failure modes compared to the same analysis performed on an AV system, and comparison between the RPNs of the two analyses could not determine which system is safer.

Variants of FMEA take a more quantitative approach by using failure rates, losses in terms of dollar values, and probability of detection instead of ordinal rankings for occurrence, severity, and detection, so that their multiplication provides an expected dollar-value loss, rather than an RPN. This approach is recommended over the RPN system for making comparisons between dissimilar systems, like AVs and human-driven vehicles.

However, even the quantitative variant of FMEA is not suitable for

analysis of AV systems due to the underlying assumption that only one failure mode occurs at any one time. Due to this assumption, FMEA does not sufficiently capture the interactions between the closely interdependent subsystems of an AV. For example, consider the case of assigning severity ratings to the failures of the various sensors on an AV. The sensor array on an AV is designed with redundancy for reliability, and it is expected that certain sensors will be impaired under certain environmental conditions (e.g. LIDAR during heavy rain or snow, vision-based cameras at night). So how should the analyst assess the severity rating of a sensor’s failure mode, when that failure will have very few consequences if other sensors remain active, but catastrophic consequences if other sensors also fail? The issue is further complicated by the response of the controller to low-confidence data. If the controller is able to recognize sensor failure and able to safely pull over or alert the driver to engage manual controls, then the danger can be averted. In this case and many others, the severity of a sensor failure depends on the response of each other sensor as well as the controller. The analysis in this thesis compensated for this by designing the rating scale (ref. Figure 4.2) to describe failures which are dependent on failures of redundant system, but this does not fully capture interdependencies in the system compared to approaches such as FTA. Like other analytical methods, FMEA fails to deconstruct the AV’s machine learning algorithms, whose function is highly dependent on the state of the overall traffic system.

Lastly, as a bottom-up approach, FMEA may be prone to incomplete lists of failure modes, especially in new technologies like AVs where all failure

modes may not be immediately apparent. As the approach depends on the analyst exhaustively listing every possible failure mode, in contrast to top-down methods which provide a method to logically reach all failure modes by starting with the entire system and reducing it to its base parts, it may be difficult for the analyst to determine when the analysis is complete.

Due to its low cost to perform, FMEA may be a useful exercise in early stages of AV design in order to organize failure modes, focus safety efforts, and develop safety measures such as schedules for routine inspections. However, its limitations preclude it from the analysis of complex dynamic systems such as AVs in real traffic. Analyzing each component of a system separately discards valuable information about how those components and their respective failure modes interact. The use of RPNs provide a measure of when a system becomes safer, but the inability to compare RPNs between systems prevents judgments on whether the safety of one system (e.g. an AV) exceeds that of another (e.g. a conventional vehicle). Performing FMEA quantitatively using failure rate data improves its ability to make meaningful comparisons, but significantly increases the cost of analysis due to the data required.

5.2.2 Fault Tree Analysis

The FTA arranged in Figure 4.6 and Table 4.3 shows how the probability of collision over the course of the left turn maneuver can be calculated from known failure probabilities for all basic events. The collision rate was estimated at 1.35×10^{-5} , or approximately 1 out of every 74,000 left turns.

Note that the probabilities for basic events were estimated for the sake of demonstration.

FTA is most often performed quantitatively, as shown in this work, such that its output (probability of failure, or the inverse probability, reliability) can be compared between dissimilar systems over a given period. While this analysis focused on the left-turn maneuver, failure rates might instead be calculated per mile of typical driving to facilitate a comparison between AVs and human-driven vehicles.

The FTA performed in this work struggled to represent some failure modes. Some failure events are not strictly binary and are not suited for Boolean logic. For example, bias error in the sensor readings is acceptable, as long as the error is less than the amount of buffer distance the controller determines is necessary between the AVs position as time of crossing the lane and the expected positions of oncoming vehicles. Similarly, the controller may initiate a turn with poor timing without causing a collision, as long as the timing is not premature or delayed enough to result in distance error exceeding the buffer between vehicles. However, collision may also occur when the sensor error and controller timing error are both within acceptable bounds if the sum of those errors exceeds the acceptable limit. FTA is intended for events that can be represented as Boolean rather than scalar variables, and may struggle to analyze a dynamic system where failure may occur due to the accumulation of acceptable errors. However, FTA could be combined with a separate error analysis to represent these interactions as Boolean events.

Additionally, for a dynamic system as complex as a traffic environment, each failure event must be very rigidly defined. For example, the failure event labeled “poor visibility” cannot simply be represented by the percent of time affected by night, thick fog, or heavy rain. It instead must refer to the probability that over a given period of time, poor visibility due to darkness or weather causes the camera sensor to fail to identify an obstacle that *would certainly* result in collision if not for other redundant systems. This data is significantly more difficult to obtain, and depends on the user’s location, usage times, and usage patterns. Therefore while the cost of FTA in terms of time and resources required to build the tree structure is low, there is significant cost associated with gathering the necessary highly-specific data regarding failure probability of basic events, if that data is not already available to the AV manufacturer.

The accuracy of the analysis is dependent on the accuracy of the data upon which it is built. Data for some events, such as the reliability of sensors, is likely to be readily available to manufacturers. Other probabilities, like the responses of other drivers, is highly dependent on the environment and state of the overall traffic system, and must be estimated. Failure probabilities for some components, such as object identification, path planning, and decision making functions, can only be obtained by other methods described in this work, such as simulation or real-world testing. Data obtained from testing in simulation is subject to the limitations previously discussed in this chapter - that is, it is not possible to fully simulate an environment with human drivers. Therefore, FTA is subject to the same limitation. Data obtained from real-

world testing introduces issues as well, since it is difficult to confirm that the driving conditions during data collection are representative of customers' driving patterns. Additionally, if enough simulation or real-world testing data is available to measure the failure rate of the AV controller, then that same data would show the failure rate of the AV system for any cause. In short, to be useful for AV analysis, FTA requires data that is most readily obtained through methods that make FTA redundant.

FTA can also be used in reverse to establish the required reliability of a system component when the desired reliability of the entire system is known. For example, if the goal is for the reliability of an AV system to be at least equal to that of a human-driven vehicle, then the desired system failure rate can be found from traffic collision data. The required reliability of the AV controller, and of each controller subroutine, can then be calculated.

Ultimately, FTA's value to AV safety assessment is dependent both on the availability of highly specific data and on whether it can be paired with a secondary method that can assess the reliability of the AV controller's machine learning algorithms that govern object identification, path planning, and decision making. *If* all data for the failure probability of basic events is available and accurate, FTA is sufficient to demonstrate how the reliability of an AV system compares to that of a human-driven vehicle; however, there is likely to be significant costs associated with gathering the necessary data. If all required data is not accurate and available, FTA is at minimum a tool for calculating the required reliability of components which have not yet been

tested.

5.2.3 Systems Theoretic Process Analysis

The STPA analysis in this work identified failure modes in finer detail than other analytical methods by framing the system in the STAMP model, and focusing on the control actions that can lead to hazards rather than the possible failure modes of each subsystem.

STPA is a purely qualitative method that does not quantify risk, but like FMEA, provides a systematic method for identifying hazards so that additional safety measures can be implemented. The end result is a list of recommended actions which ensure safety constraints are maintained. In the case of an AV performing a left turn maneuver, a sample of these actions is shown in Table 4.7. In general, the recommended actions for this analysis are secondary, qualitative analyses which can be used to guarantee the proper function of other components in the control structure. Notably, this requires a method for assessing risk in the controller’s machine learning algorithms. Although STPA alone cannot quantify safety, it can systematically list and arrange the analyses necessary for verifying safe operation.

Although STPA better accounts for the behavior of highly dynamic systems, it has a relatively high cost to perform compared to other qualitative analytical methods. The analysis in this work considered only a left turn maneuver, and generated 20 UCAs, with each UCA requiring additional analysis to identify possible causes in the system’s control structure. Extending this

analysis to all possible functions of an AV may not be feasible without generalizing beyond the point of usefulness, due to the nearly infinite conditions in a traffic environment under which a particular control action by the AV may be hazardous. For this reason, STPA may be best applied to cases which have been identified as dangerous or difficult for the AV, such as the left turn maneuver, rather than attempting to verify safety of the AV's general use, or of each possible maneuver.

Identifying failure modes exhaustively is a challenge common across analytical methods. In bottom-up methods like FMEA, it can be difficult to determine when the analysis is complete, since the analyst must list directly the system's subcomponents and failure modes as completely as possible. While top-down methods such as FTA provide a more systematic way of decomposing a system or event into its base parts, the systems-based STAMP approach used by STPA offers a distinct and unique method for identifying failure modes. Therefore it may be useful to apply STPA in conjunction with a second method, to ensure all failure modes are accounted for.

In sum, the systems-based approach of STPA is better suited for identifying the failure modes of a complex dynamic system like an AV in a traffic environment than the classical analytical methods discussed in this work. As a qualitative analysis, it is limited in making meaningful comparisons between systems; however, STPA can help structure the secondary, quantitative analyses necessary to calculate the failure rate of an AV system. Beyond quantifying risk, STPA has significant value to designers in identifying and mitigating pos-

sible errors in an AV controller.

5.3 Assessing Safety of Machine Learning Systems

A recurring criticism throughout this chapter is that many analytical methods are unable to deconstruct the machine learning algorithms used in AV object identification, path planning, and decision making, and therefore offer very limited insight into the safety of those systems. Unlike conventional software which performs static and linear procedures and demonstrates repeatable relationships between inputs and outputs, machine learning and neural network techniques often implemented in AV controllers are not straightforward to analyze and the predictions they make carry inherent uncertainty. In general, machine learning operates by reducing an input to measured quantities, called features, and assigning it a value or category, called a label, by comparison to a set of training data with known features and labels. There are a variety of algorithms for predicting the label of an example, which may compare the unlabeled example to the nearest labeled examples, or may attempt to draw boundaries through the training data to label regions. In some algorithms, the uncertainty of the prediction can be estimated statistically [56], while others do not offer an estimate of uncertainty. Regardless of the specific algorithm used, machine learning can struggle to predict cases that are rare, unexpected, or exist at the boundary of regions in the training data.

Machine learning safety is often discussed in terms of empirical risk minimization. Given features $X \in \mathcal{X}$, labels $Y \in \mathcal{Y}$, probability density function

$f_{X,Y}(x, y)$, function mapping $h \in \mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$, and loss function $L : \mathcal{Y} \times \mathcal{Y} \rightarrow R$, the risk is given as [66]:

$$E[L(h(X), Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(x), y) f_{X,Y}(x, y) dy dx$$

Because the probability density function is not known, but instead estimated by a set of m training data points, the empirical risk is as follows, where R_m^{emp} approaches R as m approaches infinity [66].

$$R_m^{emp}(h) = \frac{1}{m} \sum_{i=1}^m L(h(x_i), y_i)$$

However, in a 2016 paper formalizing the definition of safety in a machine learning context, Varshney argues that empirical risk minimization alone is insufficient, as it does not encompass various forms of epistemic uncertainty, such as cases where the training data is not representative of the true probability density function, or the probability density is particularly low in a certain region so that no training data is present there, even if the data represents the probability density function [66]. These examples are particularly applicable to the analysis of AVs, especially since the number of features required to describe and interpret a driving environment results in an exceptionally large multidimensional space. Training data might be gathered in a different environment than the vehicle is ultimately used, and the AV might encounter conditions so rare they were not represented in the training data.

The field of safety in machine learning, especially as applied to AV technology, is very new, and research into how these forms of uncertainty

can best be assessed and minimized is ongoing [17]. While better methods for assessing the safety of machine learning algorithms may be developed in the future, currently the best approach is to ensure that the training data is representative of all environments in which the AV may be used, ensure that the training data set is sufficiently large to capture rare events and edge cases that may need to poor predictions, and design the system such that predictions with low confidence alert the driver to assume manual control (as is done in all AVs with less than full autonomy). This necessitates the expensive collection of many miles of driving data, and prevents the design of AVs with full autonomy. Better methods to assess epistemic uncertainty in AV systems, combined with analytical methods such as FTA, may allow designers to calculate the failure rate of the machine learning system. Manufacturers may consider the possibility of collision due to epistemic uncertainty acceptable if the overall system collision rate is lower than for human-driven vehicles, resulting in much faster collection of training data and the development of fully autonomous vehicles.

5.4 The Future of Regulation

The core question of this work is this: How can it be determined when an autonomous vehicle is “safe enough”? As discussed in Chapter 1, regulating bodies have, for the most part, taken a very limited role with respect to autonomous driving technologies in the early stages of their development, leaving that question to the companies developing AVs and their potential

customers. Eventually, for AV technology to become widespread, the existing regulatory framework must expand to encompass it, and federal agencies will need to provide their own answer.

In predicting how regulators will respond to AV systems, one can consider how regulators responded to the introduction of conventional automobiles, since it could not at first be known how safe human drivers would be. Although the first automobiles were invented in the late 19th century, it was not until 1966 that the National Traffic and Motor Vehicle Safety Act was passed, creating the NHTSA and granting federal oversight to establish safety standards for all cars sold in the United States [11]. The long span between the adoption of automobiles and their federal regulation may suggest that federal regulation for AVs may come later than many think, although this is difficult to predict due to changing attitudes towards regulation and the role of government over the last century.

Such a delay may be unacceptable for AVs. Currently, the NHTSA spot-checks a small number of new vehicles purchased from dealerships for compliance with FMVSS. Some argue that regulation so late in the design cycle of autonomous systems is counter-intuitively more expensive than regulatory involvement early in the process [13]. As AV safety enters the forefront of public discussion, the NHTSA may face pressure to reevaluate its approach to ensuring compliance with its standards.

The approach of regulators to other autonomous technologies may suggest how the NHTSA could approach AVs in the future. One possible com-

parison is to “autopilot” systems, or flight guidance systems (FGS), used in commercial aircraft. Although automation in flight began early in the twentieth century in the form of mechanical, analog controllers [59], modern autopilot systems are highly computerized and software-dependent, and thus provide a point of comparison for AVs.

Commercial aircraft are regulated by the Federal Aviation Administration (FAA), which publishes advisory circulars (AC) as guidance for establishing compliance with airworthiness regulations. AC 25.1309-1A System Design and Analysis is a general document that establishes methods and standards for demonstrating system safety and performing safety assessments on aircraft systems. Failure conditions are categorized as “minor”, “major”, or “catastrophic”. Catastrophic failure conditions, or those involving possible fatalities, must be shown to be “extremely improbable” (having a probability on the order of 1×10^{-9} for aircraft systems, although the threshold would likely be higher for ground vehicles). The AC requires a safety assessment for catastrophic failure conditions, recommended to consist of a combination of qualitative and quantitative methods. FMEA, FTA, and reliability block diagrams (a method similar to FTA, but not discussed in this work) are the methods recommended for the safety assessment [20].

A second document, AC 25.1329-1C Approval of Flight Guidance Systems, provides more specific instructions regarding FGS. This AC gives qualitative guidelines regarding the performance of the FGS, such as what conditions it must perform under, requirements for interfaces, alerts, and con-

trols, and guidelines for integration with other aircraft systems. Along with the safety assessment described above, additional justification is required, as quoted here [21]:

“A safety assessment should be performed to identify the failure conditions, classify their hazard level according to the guidance of AC 25.1309-1A, and establish that the failure conditions occur with a probability corresponding to the hazard classification or are mitigated as intended. The safety assessment should include the rationale and coverage of the FGS protection and monitoring philosophies employed. The safety assessment should include an appropriate evaluation of each of the identified FGS failure conditions and an analysis of the exposure to common mode/cause or cascade failure in accordance with AC 25.1309-1A. Additionally, the safety assessment should include justification and description of any functional partitioning schemes employed to reduce the effect/likelihood of failures of integrated components or functions.”

Applying this approach to AV technology would place the burden on the manufacturer to convince the regulatory body that the design is appropriately safe, through any method necessary. As current methods to verify safety require extensive data that can be costly to obtain, companies might respond to this obligation in several ways: by investing in research into simulation technology, analytical methods, or safety of machine learning systems; by sharing or selling data gathered from public roads between companies; or by lobbying for the adoption of industry standards or qualitative analysis in

place of quantitative analysis.

Another challenge for future regulators is how to assess the safety of a new model or year of AV. If the process of securing NHTSA approval is very costly, it may be difficult to substantiate minor changes made year-to-year. However, if the NHTSA takes a similar approach as the FAA, AV manufacturers may be able to reuse safety assessments from past years and models to defend their new designs, as long as they are able to argue the changes are sufficiently minor or act to improve safety.

It is possible that meaningful federal regulation will not arrive until after AV technology is widespread, meaning that consumer confidence, not statistical rigor, may ultimately determine when AVs are safe enough. In this case, manufacturers will be able to use traffic collision data and sensor data gathered from their sold vehicles to facilitate any safety assessments made to the NHTSA. In the meantime, public perception would act as a check on manufacturers, who would likely compete to report the safest statistics.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis discussed five methods that have been used to assess the safety of autonomous vehicles: real-world testing, simulation, Failure Modes and Effects Analysis (FMEA), Fault Tree Analysis (FTA), and Systems Theoretic Process Analysis (STPA). Each method was shown to have advantages and disadvantages.

Though testing on public roads may be the most straightforward method, the relative rarity of fatal collisions in human driving necessitates at least hundreds of millions of miles of AV driving data to prove safety statistically, leading to unacceptable costs and delays. Simulation greatly accelerates this data collection, but it is difficult to verify that the simulated environments are representative of real environments, and it is impossible to represent other human drivers in the simulation, since the existence of a provably human-like controller would make testing of the AV controller unnecessary.

Analytical methods show some promise, but have major limitations. FMEA provides a way to systematically identify and mitigate risks, but it is based on the assumption that multiple failure modes do not occur concurrently, and therefore cannot represent the interdependent subsystems of an

AV. FTA is able to compare reliability/failure rates (over a given period of time, or over the course of an event) between systems, but only if extensive and accurate data is available for every component, including the machine learning algorithms used in object identification, path planning, and decision making. STPA accounts for the complex interactions of a dynamic system better than the older, traditional methods of hazard analysis, but does not offer a method for quantitatively comparing the reliability of two systems.

Of the methods discussed and currently available, it is the author's opinion that simulation using data from real environments is currently the most effective method. It provides valuable data regarding the rare conditions that can cause poor predictions in machine learning models, and tests the performance of the AV among real human drivers - although the behavior of human drivers in response to the AV is not reliable. While gathering data makes the process slower and more costly than using fully-simulated environments, it remains both faster and safer than testing on public roads.

As an analytical method, FTA has considerable potential to prove the reliability of AV systems, if paired with a hypothetical method for assessing reliability of machine learning algorithms. Although there is the potential for significant costs associated with gathering data for component reliability, the savings compared with reduced simulation and real-world testing could be tremendous. The implementation of this method is in part dependent on the advancement of analytical methods for assessing reliability in machine learning algorithms.

No perfect method for rigorously proving the reliability of autonomous vehicles currently exists, and it is unclear how regulators will approach the technology as it develops. Comparison to the FAA’s approach to autopilot systems in commercial aircraft suggests that AV manufacturers will be required to submit some combination of qualitative and quantitative analyses - potentially at great cost. If the current period of federal leniency continues until after the widespread adoption of AVs, data gathered from vehicles sold to the public could make these safety assessments much simpler.

6.2 Future Work

AV technology remains a very young field, and there is much work left to do in assessing their safety. Assessing the safety of machine learning algorithms in particular is a new area where more research is needed. Additionally, developments in simulation technology and changes in state and federal laws will have implications for the future of assessing AV safety.

As an extension of the work done in this thesis, the FTA performed here could be extended and compared to that of a human-driven vehicle, with more accurate values for the probability of basic failure rates, to further explore the viability of the method. An evaluation of how much of the data required for a full FTA is available, either to the public or to AV manufacturers, would also help to establish whether FTA could be used in the future to establish the safety of AV systems.

6.3 Summary

With the potential to save upwards of a million lives annually, AV technology is poised to dramatically change transportation in the coming decades. However, the scale of automotive travel combined with the relative speed with which AV systems will be rolled out poses significant risks if the safety of the vehicles cannot be verified rigorously. As automotive manufacturers compete to be among the first to deliver AVs to market, agencies have largely opted not to regulate the industry. Additionally, verifying safety statistically via testing on roads may take many decades or more, further motivating the need for alternative methods of safety assessment.

This thesis discussed five methods that have been used to assess the safety of autonomous vehicles: real-world testing, simulation, Failure Modes and Effects Analysis (FMEA), Fault Tree Analysis (FTA), and Systems Theoretic Process Analysis (STPA). By performing the latter four of these methods on a hypothetical AV system performing an unprotected left turn at an intersection, this work examined the relative ability of each approach to quantify and assess the safety of AVs. Real-world testing provides a direct approach to verifying safety, but requires prohibitively large data sets, and raises ethical concerns by exposing the public to unknown risks. Testing in simulation avoids these issues, but fully simulated environments cannot be verified to be representative of real environments. Of the analytical methods, FMEA struggles both to represent the interacting components of a system as complex as an AV, and to meaningfully quantify safety. FTA better addresses interac-

tions between components and quantifies system reliability, but is dependent on data that is difficult to obtain except by extensive real-world testing or simulation. Like FMEA, STPA does not quantify safety in a way that can be used to compare the relative safety of two systems; however, it is much better equipped to identify hazards in complex systems such as AVs, and can be combined with secondary analyses to quantify probability of the hazards identified, making it valuable as a qualitative analysis.

A recurring challenge in analytical safety analysis of AVs is characterizing the hazards associated with the controller’s machine learning algorithms, which can be far more difficult to predict than conventional software. Combining a method such as FTA or STPA with a secondary analysis capable of quantifying the risk that the controller encounters a scenario it is unable to classify with confidence may be the most promising analytical approach to assessing AV safety. However, safety in a machine learning context is a new area in which more work is needed. In the absence of an analytical approach, the most effective method currently available for assessing safety in AVs is simulation using data from real environments. This approach can be performed more quickly and safely than real-world testing, and results in a far more representative environment than full simulation.

Bibliography

- [1] 2018 Chrysler Pacifica Hybrid features & specs. <https://www.edmunds.com/chrysler/pacifica-hybrid/2018/features-specs/>. Accessed: 2019-03-1.
- [2] Automated vehicles for safety. National Highway Traffic Safety Administration. <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety#issue-road-self-driving>. Accessed: 2019-01-22.
- [3] J3016_201401 Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Report, SAE International, 2014.
- [4] Labor force statistics from the current population survey. Report, Bureau of Labor Statistics - United States Department of Labor, 2018.
- [5] Road traffic deaths data by country. Report, World Health Organization, 2018.
- [6] Autonomous vehicles/self-driving vehicles enacted legislation. Report, National Conference of State Legislatures, 3 2019.
- [7] ‘Phantom auto’ will tour city. *The Milwaukee Sentinel*, page 14, Dec. 8 1926.

- [8] E. Ackerman. Self-driving cars were just around the corner - in 1960. *IEEE Spectrum*, 2016.
- [9] American Association of State Highway and Transportation Officials. *A Policy on Design Standards - Interstate System*, 5 2016.
- [10] K. Bimbraw. Autonomous cars: Past, present and future - a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology. *2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 01:191–198, 2015.
- [11] G. C. Blomquist. *The Regulation of Motor Vehicle and Traffic Safety*. Kluwer Academic Publishers, 8 1988.
- [12] N. E. Boudette. Autopilot cited in death of Chinese Tesla driver. *The New York Times*, 9 2016.
- [13] M. L. Cummings and D. Britton. Regulating safety-critical autonomous systems: Past, present, and future perspectives. 2018.
- [14] A. Davies. Americans can’t have Audi’s super capable self-driving system. *WIRED*, 2018.
- [15] E. D. Dickmanns and A. Zapp. Autonomous high speed road vehicle guidance by computer vision. *IFAC Proceedings Volumes*, 20:221–226, 1987.

- [16] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 7 2014.
- [17] J. Duchi, P. Glynn, and R. Johari. Uncertainty on uncertainty, robustness, and simulation. <https://aicenter.stanford.edu/uncertainty-on-uncertainty-robustness-and-simulation/>. Accessed: 2019-04-7.
- [18] A. Efrati. Waymo’s big ambitions slowed by tech trouble. *The Information*, 8 2018.
- [19] D. J. Fagnant and K. Kockelman. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167–181, 2015.
- [20] Federal Aviation Administration. *AC 25.1309-1A: System Design and Analysis*, 6 1988.
- [21] Federal Aviation Administration. *AC 25.1329-1C: Approval of Flight Guidance Systems*, 10 2014.
- [22] M. E. France. Engineering for humans: A new extension to STPA. Master’s thesis, Massachusetts Institute of Technology, 6 2017.
- [23] W. Gilchrist. Modelling failure modes and effects analysis. *International Journal of Quality Reliability Management*, 10(5), 1993.

- [24] T. Griggs and D. Wakabayashi. How a self-driving uber killed a pedestrian in Arizona. *The New York Times*, 3 2018.
- [25] A. J. Hawkins. Self-driving cars continue to face little resistance from the federal government. *The Verge*, 3 2018.
- [26] M. Herger. Update: Disengagement reports 2018 - final results. *The Last Driver License Holder*, 2 2019.
- [27] J. Hughes. Car autonomy levels explained. *The Drive*, 2017.
- [28] N. Kalra and S. M. Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182–193, 2016.
- [29] T. Kanade, C. Thorpe, and W. Whittaker. Autonomous land vehicle project at CMU. *Proceedings of the 1986 ACM fourteenth annual conference on Computer science*, pages 71–80, 1986.
- [30] S. Khaiyum, B. Pal, and Y. S. Kumaraswamy. *An Approach to Utilize FMEA for Autonomous Vehicles to Forecast Decision Outcome*, volume 327, pages 701–709. Springer, Cham, 2015.
- [31] W. Knight. Waymo’s cars drive 10 million miles a day in a perilous virtual world. *Technology Review*, 10 2018.
- [32] K. Kokalitcheva. People cause most California autonomous vehicle accidents. *Axios*, 2018.

- [33] W. Kowert. The foreseeability of human-artificial intelligence interactions. *Texas Law Review*, 96(1):181–204, 2017.
- [34] J. Krafcik. “Waymo has self-driven 8 million miles on public roads, now at a rate of 25k miles per day. This real-world experience, plus over 5 billion miles in simulation, is how we’re building the worlds most experienced driver.”. Twitter. Accessed: 2019-04-09.
- [35] J. Krafcik. Waymo One: The next step on our self-driving journey. <https://medium.com/waymo/waymo-one-the-next-step-on-our-self-driving-journey-6d0c075b0e9b>. Accessed: 2019-04-8.
- [36] W. W. LaMorte. Confidence interval for a proportion in one sample. Boston University School of Public Health, 8 2016.
- [37] T. B. Lee. Fully driverless Waymo taxis are due out this year, alarming critics. *Ars Technica*, 10 2018.
- [38] W. S. Lee, D. L. Grosh, F. A. Tillman, and C. H. Lie. Fault tree analysis, methods, and applications - a review. *IEEE Transactions on Reliability*, R-34(3):194–203, 8 1985.
- [39] N. G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press, 2011.
- [40] T. Litman. Autonomous vehicle implementation predictions: Implications for transport planning. Report, Victoria Transport Policy Institute, 2018.

- [41] A. C. Madrigal. Inside Waymo’s secret world for training self-driving cars. *The Atlantic*, 8 2017.
- [42] A. C. Madrigal. 7 arguments against the autonomous-vehicle utopia. *The Atlantic*, 12 2018.
- [43] A. Marshall. We’ve been talking about self-driving car safety all wrong. *WIRED*, 10 2018.
- [44] N. Merat and A. H. Jamson. How do drivers behave in a highly automated car. In *Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pages 514–521, 6 2009.
- [45] Metamoto. Simulation as a service is now available. <https://www.metamoto.com/>. Accessed: 2019-04-29.
- [46] A. Millard-Ball. The autonomous vehicle parking problem. *Transport policy*, 75:99–108, 3 2019.
- [47] NVIDIA. NVIDIA DRIVE Constellation now available - virtual proving ground for validating autonomous vehicles. <https://nvidianews.nvidia.com/news/nvidia-drive-constellation-now-available-virtual-proving-ground-for-validating-autonomous-vehicles>. Accessed: 2019-04-15.
- [48] State of California Department of Motor Vehicles. Autonomous vehicle disengagement reports 2018. https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/disengagement_report_2018. Accessed: 2019-04-09.

- [49] State of California Department of Motor Vehicles. Report of traffic collision involving an autonomous vehicle (ol 316). https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/autonomousveh_ol316+. Accessed: 2019-04-08.
- [50] E. M. Peters, B. Burraston, and C. K. Mertz. An emotion-based model of risk perception and stigma susceptibility: Cognitive appraisals of emotion, affective reactivity, worldviews, and risk perceptions in the generation of technological stigma. *Risk Analysis*, 24(5):1349–1367, 2004.
- [51] J. Petit and S. E. Shladover. Potential cyberattacks on automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):546–556, 2015.
- [52] D. A. Pomerleau. *Neural network perception for mobile robot guidance*. PhD thesis, Carnegie Mellon University, Kluwer Academic Publishers, 101 Philip Drive, Norwell, MA 02061 USA, 1993.
- [53] Associated Press. Google’s self-driving-car project becomes a separate company: Waymo. *Los Angeles Times*, 12 2016.
- [54] G. Rogers. A primer on federal motor vehicle safety standards and avs. *Eno Transportation Weekly*, 10 2017.
- [55] B. Schoettle and M. Sivak. A preliminary analysis of real-world crashes involving self-driving vehicles. 2015.

- [56] M. Seeger. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):69–106, 2004.
- [57] D. Silver. Simulation becomes increasingly important for self-driving cars. *Forbes*, 11 2018.
- [58] R. Sparrow and M. Howard. When human beings are like drunk robots: driverless vehicles, ethics, and the future of transport. *Transportation Research Part C*, 5 2017.
- [59] L. B. Sperry. Automatic pilot for aeroplanes. Patent, 5 1922.
- [60] The Week Staff. When will self-driving cars take over? *The Week*, 2018.
- [61] D. H. Stamatis. *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. ASQ Quality Press, 2 edition, 01 2003.
- [62] D. Tokody, I. J. Mezei, and G. Schuster. *An Overview of Autonomous Intelligent Vehicle Systems*, pages 287–307. Springer, Cham, 03 2017.
- [63] J. R. Treat, N. S. Tumbas, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Castellan. Tri-level study of the causes of traffic accidents: final report. Report, Transportation Research Institute, 1979.
- [64] United States Department of Defense. *Procedures for performing a failure mode effect and critical analysis*, 11 1949.

- [65] Rowan University. Final report: Risk analysis of autonomous vehicles in mixed traffic streams. Report, University Transportation Research Center - Region 2, 5 2017.
- [66] K. R. Varshney and H. Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *CoRR*, abs/1610.01256, 2016.
- [67] A. W. Here come the self-driving taxis. *The Economist*, 11 2018.
- [68] W. Wachenfeld and H. Winner. *The Release of Autonomous Vehicles.*, pages 425–449. Springer, Berlin, Heidelberg, 05 2016.
- [69] J. Walker. The self-driving car timeline - predictions from the top 11 global automakers. *Emerj*, 2019.